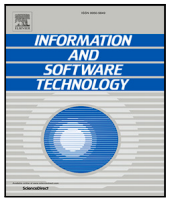
Information and Software Technology 149 (2022) 106944

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



The state of the art in measurement-based experiments on the mobile web

Omar de Munk [a], Gian Luca Scoccia [b], Ivano Malavolta [a,*]

[a] Vrije Universiteit Amsterdam, The Netherlands
[b] DISIM, University of L'Aquila, Italy

ARTICLE INFO

Keywords:
Measurement-based experiment
Mobile web
Systematic mapping study

ABSTRACT

Context: Nowadays the majority of all worldwide Web traffic comes from mobile devices, as we tend to primarily rely on the browsers installed on our smartphones and tablets (e.g., Chrome for Android, Safari for iOS) for accessing online services. A market of such a large scale leads to an extremely fierce competition, where it is of paramount importance that the developed mobile Web apps are of high quality, e.g., in terms of performance, energy consumption, security, usability. In order to objectively assess the quality of mobile Web apps, practitioners and researchers are conducting experiments based on the measurement of run-time metrics such as battery discharge, CPU and memory usage, number and type of network requests, etc.
Objective: The objective of this work is to identify, classify, and evaluate the state of the art of conducting measurement-based experiments on the mobile Web. Specifically, we focus on (i) which metrics are employed during experimentation, how they are measured, and how they are analyzed; (ii) the platforms chosen to run the experiments; (iii) what subjects are used; (iv) the used tools and environments under which the experiments are run.
Method: We apply the systematic mapping methodology. Starting from a search process that identified 786 potentially relevant studies, we selected a set of 33 primary studies following a rigorous selection procedure. We defined and applied a classification framework to them to extract data and gather relevant insights.
Results: This work contributes with (i) a classification framework for measurement-based experiments on the mobile Web; (ii) a systematic map of current research on the topic; (iii) a discussion of emergent findings and challenges, and resulting implications for future research.
Conclusion: This study provides a rigorous and replicable map of the state of the art of conducting measurement-based experiments on the mobile Web. Its results can benefit researchers and practitioners by presenting common techniques, empirical practices, and tools to properly conduct measurement-based experiments on the mobile Web.

1. Introduction

The mobile Web is everyday increasingly important, as more and more people rely primarily on a mobile device to access online services. In November 2020, more than 55% of all worldwide Web traffic came from mobile devices while, in the same month of the year 2015, this percentage was only 42% [1]. For a large group of people, their mobile device is the primary means of accessing the Internet.

Alongside this growth, mobile browsers (e.g., Chrome for Android, Safari for iOS) are evolving into a fully-featured complex software platform, thanks to the continuous development of the HTML5 specification and to the constant addition of newer APIs that provide a bridge to interact with hardware sensors and novel software capabilities, *e.g.,* geolocation, motion sensors, and speech recognition [2]. The expectations of users in terms of quality have increased drastically when browsing the Web on their mobile device and thus are more relevant than ever. Various sources have shown the impact of quality-related aspects of mobile Web apps like performance, usability, and security in terms of revenue and user retention. As an example, Amazon calculated that a page load slowdown of just one second could cost them $1.6 billion in sales each year [3].

The quality of a mobile Web app can be affected by a wide variety of internal (*e.g.,* page parsing and rendering speed) and external (*e.g.,* mobile network conditions) factors, which can also interact with each other in unexpected ways. Hence, improving and even just assessing the overall quality of mobile Web apps is a complex and challenging task [4]. As a result, a growing number of studies is investigating quality-related aspects of mobile Web apps by conducting *measurement based experiments*. In this study we refer to measurement-based experiments as those experiments whose dependent/independent variables

* Corresponding author.
E-mail addresses: o.de.munk@student.vu.nl (O. de Munk), gianluca.scoccia@univaq.it (G.L. Scoccia), i.malavolta@vu.nl (I. Malavolta).

https://doi.org/10.1016/j.infsof.2022.106944
Received 29 November 2021; Received in revised form 4 May 2022; Accepted 11 May 2022
Available online 20 May 2022

are based on measures collected at run-time, such as battery discharge, CPU and memory usage, number and type of network requests, etc [4]. Examples of experiments that are not measurement-based include: qualitative studies based on developers' interviews or online questionnaires, empirical studies focussing on statically-collected metrics (*e.g.,* via program analysis), empirical studies focussing on mining software repositories, secondary studies.

The **goal** of this paper is to carry out a review of existing studies that conduct measurement-based experiments on the mobile Web by applying the systematic mapping methodology. Starting from a search process that identified 786 potentially relevant studies, we reduced it to a set of 33 primary studies following a rigorous selection procedure. We defined and applied a classification framework to them to extract data and gather insights. Finally, the obtained data is synthesized with the goal of presenting a clear overview of the state-of-the-art of conducting measurement-based experiments on the mobile Web.

The main **contributions** of this study include:

- a reusable framework for classifying measurement-based experiments on the mobile Web in terms of their used metrics and data management strategies, platforms, subjects, and execution setup;
- an up-to-date map of the state of the art in measurement-based experiments on the mobile Web;
- an evidence-based discussion of the emerging results, and their implications for future research;
- a replication package for independent verification and replication.

The **main motivations** for conducting this study are: (i) the specificity and technical challenges of carrying out measurement-based experiments on the mobile Web, (ii) the fragmentation of the publications landscape on the topic, and (iii) the lack of shared research directions for the involved research communities Specifically, mobile-specific factors have to be taken into consideration when conducting experiments on the mobile Web. For example, energy consumption plays a critical role in mobile Web apps, as mobile devices are equipped with a limited battery and often only have intermittent access to the electricity grid. Another example is bandwidth usage, as cellular networks are often slower and more expensive than their cabled counterpart. Also, conducting and reporting measurement-based experiments is a challenging task: significant experience and knowledge in a wide variety of different areas, such as empirical software engineering, statistics, programming, and networking is required. A wide range of possible questions may arise while planning measurement-based experiments. For instance, what are the more suitable metrics to quantify the phenomena of interest? Which kind of mobile device is best suited to run experiments on? Are software-based measurements tools equally as valid as hardware-based ones? How to analyze the collected data? What are the expectations with regard to the studies' replicability? We give guidance to both researchers and practitioners on these questions by rigorously analyzing existing scientific studies on the mobile Web. About the publications landscape, before conducting this study we had anecdotal indications that measurement-based experiments targeting the mobile Web are published across several scientific venues across different scientific areas, such as software engineering (*e.g.,* EASE, ICSME), mobile systems (*e.g.,* MOBICOM, MOBILESoft, HotMobile), the Web (*e.g.,* WWW, Internet Computing, ICWE), and networking (*e.g.,* Computer Networks, INFOCOM); we anticipate that this indication has been confirmed in this study. With this study we mitigate such fragmentation problem by providing a unified map to both novice and expert researchers of current research on measurement-based experiments on the mobile Web. The previously-mentioned fragmentation also caused a lack of shared research directions for future investigations. By building on the results emerging from this study, in Section 5 we fill this gap by providing an in-depth discussion of the main implications and recommendations for both researchers and practitioners in the field.

Table 1
Goal of this study.

Purpose	Identify, classify, and summarize
Issue	the characteristics of
Object	measurement-based experiments
Context	on the Mobile Web
Viewpoint	from a researcher's and practitioner's point of view.

This study is an extended version of our previous research on measurement-based experiments on the mobile Web [5]. The new contributions of this study are: (i) the extension of the set of primary studies via a new automatic search to cover publications until the end of September 2021 and backward/forward snowballing, (ii) a more in-depth elaboration of the extracted data, (iii) the analysis of the research trends over the years, and (iv) the orthogonal analysis about the potential interactions between various parameters of the classification framework.

The **target audience** for this paper includes both practitioners and researchers that are interested in conducting measurement-based experiments on the mobile Web and that want to be aware of state-of-the-art empirical practices, techniques, and tools used in such experiments.

2. Study design

In this section we present the design of this study. This study is designed and carried out by following well-accepted methodological guidelines on secondary studies [6–8]. As shown in Fig. 1, this study has been designed as a four-phases process: planning, search and selection, data extraction, and data synthesis. In the remainder of this section we will describe each of those phases.

2.1. Phase 1: Planning

The main goal of this phase is to establish the scope of the study (*i.e.,* the *why*) and to plan the activities to be carried out (*i.e.,* the *how*) [8]. Specifically, we firstly formalize the goal and research questions of the study (Section 2.1.1), and then we create a plan for carrying out all the other activities of this study (Section 2.1.2).

2.1.1. Goals and research questions definition

The goal of this study is to identify and classify the characteristics of existing research that conduct measurement-based experiments on the mobile Web. More specifically, we formulate such high-level goal by using the Goal-Question-Metric perspectives proposed by Basili et al. [9].

Table 1 shows the result of the above mentioned formulation.

To achieve the above-mentioned research goal, we ask the following research questions (RQs).

[RQ1] *Which **metrics and data management practices** are considered when conducting measurement-based experiments on the mobile Web?*
Rationale – Over the years researchers carried out a plethora of measurement-based experiments with different viewpoints, methodological approaches, and targeting specific sub-problems. For example, experiments on the mobile Web can focus on different aspects of the mobile Web (*e.g.,* energy consumption [10], performance [11], networking and caching [12]), used metrics (*e.g.,* consumed joules, page load time, cache hit rate), procedures followed for analyzing the collected measures (*e.g.,* from simple descriptive statistics to hypothesis testing and effect size estimation).

By answering this research question we support researchers by providing (i) a solid foundation for classifying existing (and future) empirical research on the mobile Web, (ii) an understanding of current research gaps of the state of the art, and (iii) a reference for looking up how specific data analysis methods have been applied in already-performed experiments. Practitioners can use the answers to

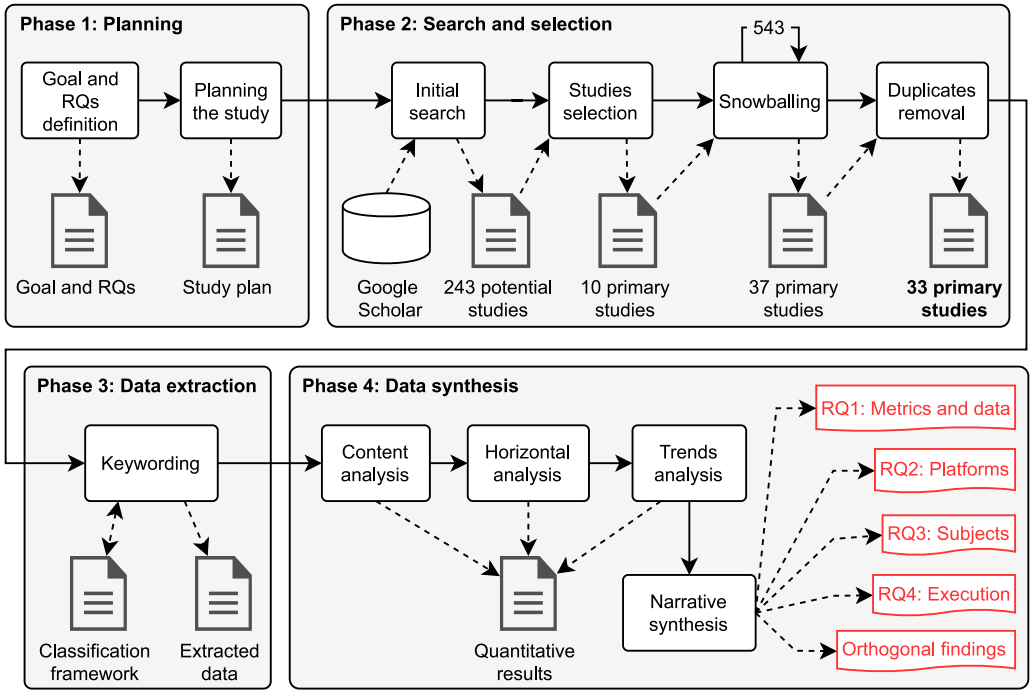


Fig. 1. Overview of the study design.

this research question for (i) identifying experiments whose results can be used in their specific projects and organizations (*e.g.,* which empirically-backed techniques can be used for saving energy) and (ii) having access to a catalog of already-used metrics that researchers used in their experiments, which can then be used internally in their own industrial projects.

[RQ2] *Which **platforms** are considered when conducting measurement-based experiments on the mobile Web?*

Rationale – In this context, with the term "platform" we mean the environment where the web apps are running, defined as the hardware device (*e.g.,* a smartphone or tablet, or an emulator), the operating system, and the browser. The choice of the platform where an experiment is executed can strongly impact the results of the experiment itself and can even make an experiment unfeasible; for example, emulators are generally not used in experiments targeting energy consumption since energy is a physical resource which strongly depends on physical processes (*e.g.,* the temperature of the battery).

Researchers can benefit from the results of this research question since they can get a clear indication about the mostly used devices, OSs, and browsers used in state-of-the-art research on measurement-based experiments on the mobile Web. The built map has the potential to unveil research gaps and opportunities for replicating experiments in more modern/realistic platform settings.

[RQ3] *Which **subjects** are considered when conducting measurement-based experiments on the mobile Web?*

Rationale – The selection of the subjects is a fundamental aspect in any empirical study, especially for its external validity [8]. In the context of experiments on the mobile Web, the choice of the subjects generally boils down to choosing a large-enough sample of mobile Web apps which are representative of the targeted population (*e.g.,* Progressive Web Apps [13]). The choice of the subjects of experiments on the mobile Web apps is currently more an art than a fixed procedure, where researchers must consider several decision points, such as whether the subjects should be real or synthetic Web apps, the sources from which subjects are sampled (*e.g.,* the famous Alexa list of the top 1M sites[1] or other more robust sources like the Tranco list [14]), the number of subjects (which might impact the feasibility of the experiment), etc.

Our answer to this research question will support researchers and practitioners in weighting the validity of the experiments carried out until now and in comparing the number and types of subjects of their own experiments against the state of the art. For example, if a researcher will need to carry out an experiment involving a large number of real in-production Web apps and to host them locally in their experimental infrastructure, then they can use our extracted map to identify and study those studies that are compatible with their setup and build on the lessons learned by the other researchers.

[RQ4] *What is the state-of-the-art on the **execution** of measurement-based experiments on the mobile Web?*

Rationale – The validity of measurement-based experiment is heavily rooted on its actual execution and on the used measurement infrastructure. This is a non-trivial and multi-faceted problem and also in this case researchers have adopted different solutions over the years. For example, it is important to define the scope of the execution of a Web app (*e.g.,* is it only loaded in the browser, or does it require to simulate users via usage scenarios?), which components of the Web app are relevant (*e.g.,* HTML, JSS, CSS), the network conditions (*e.g.,* WiFi vs 4G/5G), the status of the cache, and, last but not the least, which tools are used to carry out the measurement (*e.g.,* the well-known Monsoon hardware power monitor,[2] Google Lighthouse[3] for performance and quality audits, etc..).

Similar to the other research questions, our answer to this RQ will benefit both researchers and practitioners by providing a solid and systematic overview of the current technical choices and solutions adopted by researchers on the mobile Web. Such an overview can potentially guide researchers and practitioners in taking better informed decisions for their future experiments (instead of reinventing the wheel).

Overall, by answering these research questions we provide an overview of the possibilities, common practices, and approaches to carry out measurement-based experiments on the mobile Web. Answering these questions is useful for researchers and practitioners as

[1] https://www.alexa.com/topsites

[2] https://www.msoon.com/high-voltage-power-monitor

[3] https://developers.google.com/web/tools/lighthouse

it helps them with positioning, planning, and conducting their own experiments, while building on a solid foundation coming from the common experience of the community.

2.1.2. Planning the study

In this step we create a plan for carrying out all the other phases of this study, with a special emphasis on how each phase contributes to answering the previously-mentioned research questions. In order to mitigate potential threats to validity, the plan has been iteratively discussed among all the authors and defined *a priori*. The planned activities are described in Sections 2.2, 2.3, 2.4.

In this phase we also setup a GitHub repository[4] for the **replication package** of this study. The replication package is publicly available for independent replication and verification of our study. The replication package includes the raw data of our search and selection phase, the raw data extracted from each primary study, the extracted keywords and themes emerging from our content analysis, all contingency tables we built for the horizontal analysis, and the scripts we developed for data exploration and analysis.

2.2. Phase 2: Search and selection

The main goal of this phase is to retrieve a representative set of scientific studies reporting measurement-based experiments on the mobile Web. As shown in Fig. 1, our search and selection phase has been designed as a multi-stage process; this gives us full control on the number and characteristics of the entries being either selected or excluded during the various stages. We carried out those steps in a sequential order and independently of each others (for the sake of replicability and independent verification). In the following we provide the details about each step.

2.2.1. Initial search

In this step we execute an automated search query. The search query is executed on *Google Scholar*, which is considered to be one of the most comprehensive academic search engines currently available [15]. In addition, we use Google Scholar as data source for the following main reasons: (i) it is one of the largest and most complete databases and indexing systems for scientific literature; (ii) as reported in [16], the adoption of this data source has proved to be a sound choice to identify the initial set of literature studies for the snowballing process [16]; (iii) the query results can be automatically processed via already existing tools. The query used to the perform the automated search is provided in Listing 1 and is applied to the title of the targeted studies.

```
("Web" OR "browser") AND ("Experiment" OR "Empirical" OR
"assessment" OR "Analysis" OR "Measurement" OR "assessing" OR
"Analysing" OR "Measuring") AND ("mobile")
```

Listing 1: Search string used for the automatic search

In essence, the search string can be divided into three main components separated by the AND logical operator, of which the first one captures the focus on the Web, the second one is about measurement-based experiments, and the third one keeps the focus on the mobile domain. The automatic search is executed at the end of September 2021. This phase leads to the identification of 243 potentially-relevant studies.

[4] Replication package of this study: https://github.com/S2-group/IST-2022-replication-package.

2.2.2. Studies selection

In this step, we filter the 243 potentially-relevant studies by rigorously applying a set of inclusion and exclusion criteria. A study is added to the set of primary studies if it satisfies **all** inclusion criteria and **none** of the exclusion criteria. We used the following inclusion and exclusion criteria:

- IC1 – Studies focusing on the mobile Web.
- IC2 – Studies reporting measurement-based experiments, *i.e.,* their findings are based on quantitative data collected at run-time (*e.g.,* page load time, energy consumption, etc..).
- IC3 – Studies targeting Web apps running either on a smartphone or a tablet.
- EC1 – Studies that are not written in English.
- EC2 – Studies for which the full text is not available.
- EC3 – Secondary or tertiary studies.
- EC4 – Studies that are not in the form of a journal article, conference paper, book or book section.
- EC5 – Studies that have not been peer reviewed.
- EC6 – Studies whose main contribution is not an empirical evaluation.

Each study is manually analyzed by applying the adaptive reading depth technique [17], *i.e.,* by incrementally reading the text, starting with the title, abstract, and introduction, and then reading the full text, if necessary.

Furthermore, syntactic duplicates (papers that are exactly the same, *i.e.,* same title, authors, abstract, and venue) are excluded and thus just a single version is kept. After evaluating all 243 potentially-relevant studies, a total of 10 studies meet the inclusion and exclusion criteria.

2.2.3. Backward/forward snowballing

The main goal of this step is to complement the previously-described automatic search with a snowballing activity [16]. Snowballing allows us to enlarge the set of potentially-relevant studies by (i) considering each study selected in the previous phases and (ii) selecting those papers that are either cited by it (backward snowballing) or citing it (forward snowballing). We perform a closed recursive backward and forward snowballing activity in this study [16], *i.e.,* we iterate over all currently-considered studies and do snowballing until there is no other study to evaluate. The snowballing activity leads to 543 additional studies on which we again apply the same selection criteria used in the previous step. This round leads to the inclusion of 27 additional studies meeting our selection criteria, leading to a total of 37 primary studies.

2.2.4. Duplicates removal

During this step it is still possible for papers to be excluded from the set of primary studies. This happens if it turns out that a study is found to be a duplicate of another study. Indeed, if more than one potentially-relevant study is about the same experiment (*e.g.,* a conference paper that is extended to a journal version), only one instance is considered. We identified two pairs of studies (S4 and S31) that have two publications about the same experiment, and one experiment (S9) that iss spread across three different publications. After merging these papers into a single entry, we obtain a final number of 33 primary studies. Generally, the journal version is preferred, since more complete, but both versions are used in the data extraction phase and in the trends analysis. This is necessary for ensuring completeness and traceability of the obtained results [8].

2.3. Phase 3: Data extraction

The main goal of this phase is to collect from the primary studies the relevant information to answer our research questions. In this phase we manually collect data from each primary study. The data extraction phase is performed collaboratively by two of the authors of this study.

Table 2
The classification framework.

Parameter	*Definition*	*Possible values*
Metrics and data management (RQ1)		
Main aspect	The aspects of mobile web apps targeted by the experiment.	Energy Consumption (EC), Performance (PF), Bandwidth (BW), Caching (C), or Memory Consumption (MC)
Used metrics	The metrics collected during the experiment.	Joules, Page Load Time (PLT), Bytes, Hit Rate
Data analysis	The type of data analysis carried out during the experiment.	Descriptive Statistics (DS), Correlation Analysis (CA), Development of Predictive Models (PM), Hypothesis Testing (HT), Effect Size Estimation (ESE)
Replicat. package	The artifacts present in the replication package of the experiment	Instructions, Code & Data (ICD), Code & Data (CD), Code only (C), None (NO)
Platform (RQ2)		
Device type	The devices used in the experiment.	Smartphone, Tablet, Emulator
OS	The operating system running on the mobile device during the experiment.	Android, iOS, Other
Browser	The browser used during the experiment.	Chrome, Safari, FireFox, Modified, Other
Subjects (RQ3)		
Type	Whether the subjects are real-world Web apps or apps developed for the experiment.	Both, Real, Synthetic
Selection	The source from which the real Web apps are sampled.	Alexa, No source, List, Other
Hosting	Whether the real Web apps are copied to another server (a mirror) or kept on their own original server.	Original, Mirrored
Nr. of subjects	Number of Web apps used in the experiment.	Integer
Subjects provided	Whether the paper explicitly mentions the Web apps used during the experiment.	Yes, No
Experiment execution (RQ4)		
Scope	To what extent the experiment executes each Web app.	Page load only, Usage scenarios
Focus	The components on which the experiment focusses on.	All, HTML, CSS, JS
Tools	What tools (hardware or software) are used to carry out the measurement.	Monsoon power monitor, Google Lighthouse, Custom JavaScript
Network cond.	The type(s) of network considered while executing the experiment.	WiFi, 3G, 4G, Simulated
Caching	Whether the browser cache is cleared before each run of the experiment.	Enabled, Disabled, Not reported

In order to have a rigorous data extraction process and to ease the management of the extracted data, a well-structured classification framework has been rigorously designed [7]. The resulting classification framework is shown in Table 2. We designed the comparison framework so to facilitate the search for overarching themes and patterns among the primary studies in terms of how they conduct measurement-based experiments on the mobile Web. Our classification framework is composed of four facets, each of them addressing its corresponding research question: metrics and data management (RQ1), platform (RQ2), subjects (RQ3), experiment execution (RQ4).

In order to have a rigorous data extraction process and to ease the management of the extracted data, a well-structured data extraction form will be designed upfront. The form is composed of the various parameters of the classification framework. For each primary study, a researcher collects in a spreadsheet a record with the extracted information for subsequent analysis: the spreadsheet columns will be the parameters, while each spreadsheet row will represent the data of each primary study.

For each facet of the classification framework, we follow a systematic process called *keywording* [18] for defining its main parameters and their corresponding values. The goal of the keywording process is to effectively develop a classification framework so that it fits (i) the characteristics of the primary studies and (ii) the goal and research questions of this research [18]. Specifically, in line with previous secondary studies on other topics [19,20], our keywording process is composed of the following main steps. Firstly, as also suggested in [8], we randomly select a set of four primary studies to be used as pilot studies. Then, a researcher collects keywords and concepts by reading the full-text of each pilot primary study. When all pilot primary studies have been analyzed, all keywords and concepts are combined together to clearly identify the emerging characteristics of the research on measurement-based experiment on mobile Web apps (in this step also a second researcher has been involved). The output of this step is the initial version of the classification framework. Now, for each subsequent primary studies, we (i) extract information about the study and (ii) collect any kind of additional information that is considered relevant, but does not fit within any parameter of the classification framework. If the collected information about the current primary study fits completely within the classification framework, then we proceed to analyze the next primary study, otherwise the classification framework is discussed among all two researchers and possibly refined accordingly. This process ends when all primary studies are analyzed. The specific parameters emerging from the keywording process are independent from each other and are extracted independently; they are described in details in Section 3.

2.4. Phase 4: Data synthesis

The main goal of this phase is to extract key findings from the data extracted from the primary studies in order to build the map of current research on measurement-based experiments on the mobile Web [21, § 6.5]. This phase is composed of four main steps: content analysis (see Section 2.4.1), horizontal analysis (see Section 2.4.2), trends analysis (see Section 2.4.3), and narrative synthesis (see Section 2.4.4).

2.4.1. Content analysis

In the context of this study, the goal of content analysis is to obtain a quantitative assessment of the extracted data (*e.g.,* the frequency of experiments targeting energy efficiency vs those targeting performance) [21]. To do so, depending on the specific parameter to be

analyzed, we apply descriptive statistics and create bar plots and tables for better understanding the data and the emerging patterns. The results of our content analysis are used as input to the narrative synthesis step (Section 2.4.4), whose results are then grouped according to our research questions and presented in Section 3.

2.4.2. Horizontal analysis

The main goal of the horizontal analysis is to investigate on the existence of possible interesting relations between data pertaining to different parameters of the classification framework (*e.g.,* if the aspect being investigated corresponds to experiments with a higher/lower number of subjects).

Our horizontal analysis is carried out by following the steps and lessons learned in our previous secondary studies (*e.g.,* [20,22]):

1. We automatically create a contingency table for each possible pair of parameters of the classification framework, leading to a total of 210 contingency tables; all contingency tables are available in the replication package of this study, allowing independent researchers to further investigate them.
2. Two researchers collaboratively analyze each parameter of the classification framework and create a set of 15 pairs of parameters whose relationship is deemed relevant to be investigated. For example, we create the $< Main\ aspect, Device\ type >$ pair in order to understand if real devices are preferred to emulators when investigating on specific aspects of mobile web apps (e.g., energy), or we create the $< Type, Network\ conditions >$ pair in order to understand if real or synthetic mobile web apps are considered more frequently under different network conditions, etc. All 15 potentially-relevant pairs and our notes about their analysis are available in the replication package.
3. We iteratively analyze the contingency table of each of the 15 potentially-relevant pairs and keep track of the main emerging results. To clarify how extracted the emerging results for each pair, we consider as an explanatory example the case involving the following pair of concepts: $< Main\ aspect, Number\ of\ subjects >$; their corresponding extracted data is reported in Fig. 5. based on our experience in the field, when building this pair we had the following hypotheses to test: (i) "*studies on energy consumption will tend to have a lower number of subjects due to the notoriously higher complexity and duration of experiments on energy*", "*studies on bandwidth requirements and caching will tend to have a higher number of subjects since those experiments can be easily replicated by reusing previously-recorded network traces*".
4. We iteratively analyze the contingency table of each of the 15 potentially relevant pairs and keep track of the main emerging results.
5. We filter out all the results which were either not supported by a sufficient number of data points or not revealing any evident pattern. This filtering step is performed *manually* and collaboratively by two researchers on a pair-by-pair fashion, until a full agreement about the inclusion of each pair is reached. Because of the semantic nature of the selected pairs and for avoiding false negatives, we decided to do not apply any fixed rule in this filtering step (i.e., a specific number of data points or quantitative criteria to detect evident patterns).
6. The remaining relevant contingency tables are used as input of the narrative synthesis step (Section 2.4.4). Section 4 reports the main results emerging from our horizontal analysis.

2.4.3. Trends analysis

When performing trends analysis, we focus on how each possible value of all parameters of the classification framework evolves over time. In order to do so, for each parameter we (i) create a line plot with a line for each possible value of the parameter, years in the X axis, and the number of primary studies on the Y axis,[5] (ii) collaboratively discuss the line plot, and (iii) keep track of emerging results in the form of notes. The line plots and the notes are used as input of the narrative synthesis step (Section 2.4.4). The results of the trends analysis are reported in Section 3 for each research question.

Table 3
Main aspects.

Main aspect	# Studies	Studies
Performance	19	S2, S6, S7, S8, S9, S11, S13, S14, S15, S17, S21, S22, S23, S25, S27, S28, S30, S31, S32
Energy consumption	16	S3, S4, S5, S6, S9, S14, S16, S18, S19, S20, S24, S25, S26, S29, S32, S33
Bandwidth	5	S12, S16, S19, S28, S32
Memory consumption	3	S1, S28, S32
Caching	2	S9, S12

2.4.4. Narrative synthesis

Narrative synthesis refers to the method of synthesizing research in the context of systematic reviews where a textual narrative summary is adopted to explain the characteristics of the primary studies [21,23]. In this study we take as input the quantitative data and notes emerging from the content analysis and trends analysis, and then we describe the main obtained findings on a parameter-by-parameter fashion. For the horizontal analysis, we describe the emerging findings based on the identified relevant contingency tables.

3. Results

In this section we report the insights gained from our analysis of the extracted data for each research question.

3.1. Metrics and data management (RQ1)

3.1.1. Main aspect

Table 3 shows the frequency of the main aspects across the set of primary studies. As can be observed, the two most frequently considered aspects are performance and energy consumption, respectively. Far less common are studies that examine the impact of bandwidth, cache performance and memory consumption.

Out of the 33 studies there are 11 papers that aim to quantify multiple aspects as their main focus. Not surprisingly, the combination of measuring energy consumption and performance is seen most often followed by studies focusing on energy consumption and bandwidth. They together make up more than half of this category.

Example. S15 argues that the available content on smartphones, apps, and the web, comes in two versions: (i) free content monetized via advertisements (ads); and (ii) paid content monetized by user subscription fees. The authors describe an approach that enables the separation of web contents in websites and use it to evaluate the energy cost due to downloading, rendering, and displaying web ads over Wi-Fi and 3G networks. That is, how much energy web ads consume when a user accesses the web. Their results highlight that ads on smartphones come with a high cost that must be considered by the designers and vendors of apps.

[5] It is important to note that in this step we consider all 37 primary studies before the duplicates removal step in Phase 2 (see Fig. 1).

3.1.2. Used metrics

The diversity in metrics used to quantify the aforementioned aspects differ greatly. When reporting energy consumption 10 of the 16 studies use Joules to do so. The five papers that deviate from this practice convey their measurements in milliampere-seconds (mAs) (S19), millivolt-seconds (mVs) (S31) or by comparing them to a baseline (S10, S21, S33)

The five studies measuring bandwidth usage all use bytes or a derivative thereof, e.g. kB (S19, S20, S21, S28, S31).

The cache performance, measured by 2 primary studies, is reported by using the hit rate, defined as the division of saved traffic and total traffic in a visit (S6). On a deeper level more advanced cache performance metrics are used such as the cacheability and actual cache performance of a webpage together with the positive and negative hit and miss ratios (S15).

One of the studies (S1) that report memory consumption uses the Proportional Set Size (PSS) as a metric. PPS is defined as the portion of memory occupied by a process and is composed of the private memory of that process plus the proportion of shared memory with one or more other processes. The other two studies (S25, S32) that measure memory consumption simply use megabytes (MB).

When it comes to performance we see much more variety in terms of used metrics. A total of 19 different metrics were found. However, this observed influx of different performance metrics is predominantly caused by a single study: S3. Overall, 9 of the 19 studies measuring performance use the Page Load Time (PLT) metric, followed by the SpeedIndex (SI) which is used by 4 studies and time to interactive (TTI), utilized by 2 papers. One study (S29) relies on the user-perceived Page Load Time (uPLT). We have also encountered studies that use a relatively undefined performance metric, such as browser latency (S26) and loading time (S22, S6), but do not give a solid definition and its therefore unclear how and if they actually differ from PLT.

Example. S29 investigates whether Quality of Experience (QoE) metrics designed for the desktop environment are equally effective in estimating the QoE in the mobile domain. After developing a system for collecting and analyzing mobile web experiences, they collected experimental measurements from 100 participants. The analysis of the collected user data highlights that QoE metrics designed for the desktop environment are not necessarily adequate for the mobile environment, and appropriate metrics should be devised to reflect the mobile web experience.

3.1.3. Data analysis

Table 4 lists primary studies by employed data analysis technique. As can be observed all 33 primary studies analyzed their gathered data using descriptive statistics, i.e. using mean values, standard deviations and presenting plots to get an understanding of the data that has been collected. 11 out of the 33 papers use hypothesis testing to support their findings and decisions with statistical evidence. Effect size estimation, used to measure the strength of the relationship between variables, is utilized by 6 papers. To gain insights into the relationship between variables, 3 studies make use correlation analysis techniques. Finally we found 3 studies that develop prediction models based on their gathered data.

Example. S33 analyzes the difference in energy consumption of Progressive Web Apps (PWAs), focusing on UI rendering and interaction scenarios. After implementing five versions of the same app with different development approaches, their energy footprint on two Android devices is collected during four execution scenarios. Multiple two-tailed null hypotheses and corresponding alternative hypothesis have been formulated and tested employing the Mann–Whitney U test [24] with the Bonferroni correction [25]. Finally, when relevant differences in energy consumption were detected, the effect size was calculated by applying Cliff's Delta [26].

Table 4
Data analysis techniques.

Analysis technique	# Studies	Studies
Descriptive statistics	33	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22, S23, S24, S25, S26, S27, S28, S29, S30, S31, S32, S33
Hypothesis testing	11	S4, S5, S6, S8, S13, S17, S21, S23, S24, S25
Effect size estimation	6	S5, S6, S8, S13, S25, S33
Correlation analysisn	4	S5, S6, S9, S23
Predictive models	3	S5, S23, S24

Table 5
Replication package availability.

Package availability	# Studies	Studies
None	22	S3, S4, S5, S7, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S26, S27, S28, S29, S30, S33
Instructions, code & data	6	S2, S6, S8, S24, S25, S33
Code & data	2	S21, S23
Code only	2	S22, S31
Data only	1	S32

3.1.4. Replication package

Table 5 depicts the availability of replication packages across the 33 primary studies. We observe that 22 of the 33 studies do not provide a replication package at all. For the 11 studies that actually do provide a replication package we find that both the contents and the quality of the packages differ significantly. 6 papers equip the reader with a detailed set of written instructions on how to use its contents to replicate the experiment in combination with the code and the collected data. 2 papers provide the code and data but do not elaborate on how to actually use these (S5, S8) while 2 studies (S17, S30) only provide the code which consisted of their experimental apparatus. 1 study (S31) makes the collected data available to other researchers but does not provide the used code or instructions on how to use the data.

3.1.5. Trend analysis

In this section, we report our analysis of the research trends over the years for the parameters related to RQ1.

We report the yearly *number of studies* that have conducted measurement experiments on the mobile web in the plot of Fig. 2-A. From the plot, it can be observed that the number of studies published on the topic has been growing over the years, with recent years having 4 or more papers per year, as opposed to less recent ones that had at most 2 papers up until the year 2013. The only exception is the year 2018, for which only 1 relevant paper was found. The year 2015 record the highest amount, with 7 published papers in a single year.

Concerning the *main aspect* investigated by analyzed papers, we reported the yearly breakdown in Fig. 2-B. Performance and energy consumption are the most investigated topics, as previously reported in Section 3.1.1. However, the former appears to be experiencing a growing interest in the more recent years (3 papers in 2019, 5 in 2020), while interest for the latter appears to have peaked in the year 2017 (4 papers) and has been less investigated ever since. Observing the less popular aspects, similar considerations can be made for memory consumption, that appears to have been more investigated in more recent years as opposed to less recent ones.

Fig. 2-C reports trends for the *data analysis* techniques used in the analyzed papers. As reported in Section 3.1.3 descriptive statistics are the most used analysis technique and have been used exclusively in

Table 6
Device type.

Device type	# Studies	Studies
Smartphone	25	S1, S2, S3, S4, S6, S7, S8, S10, S13, S14, S15, S17, S18, S19, S20, S23, S24, S25, S26, S28, S29, S30, S31, S32, S33
Emulation	6	S10, S12, S13, S17, S23, S28
Tablet	4	S6, S11, S15, S22

the years up to 2017. From this year onward, the adoption of other analysis techniques has experienced a growing trend. We consider this trend positive as the usage of more robust analysis techniques increases the likelihood of reporting valid results.

Finally, the yearly breakdown for the availability of the *replication package* is visualized in Fig. 2-D. The majority of papers do not make available either a partial or complete replication package. However, the number of studies that do make it available has been increasing in the more recent years, with the first one appearing in 2015 and the maximum being recorded in the year 2020 with five papers. This trend is encouraging as having an available replication package helps in making the obtained results more credible, reproducible, and replicable by the community.

Summary of the main findings (RQ1):

- The most investigated aspects of mobile Web apps are performance (also growing over time) and energy consumption, followed by network-related aspects such as bandwidth and caching.
- The landscape of the used metrics is extremely fragmented.
- Descriptive statistics are the most used data analysis procedure; recently, also hypothesis testing, effect size estimation, and correlation analysis are carried out by researchers.
- A minority of studies provide a replication package for independent verification and replication of the study; when present (mostly recently), the replication package tends to be complete (*i.e.,* it contains both instructions, developed code, and raw data).

3.2. Platform (RQ2)

3.2.1. Device type

The frequency of device types on which the experiments are carried out is given in Table 6. We can see that most studies, 25 of the 33 papers, run their experiments on a smartphone. Most of these smartphones are phones such as these that can be found in the Google Nexus or Samsung Galaxy product line. An interesting exception is a study (S14) that uses an Odroid-XU3 board that contains an Exynos5422 SoC which is also used in the Samsung Galaxy S5 phone. It runs the Android 4 KitKat OS and therefore essentially functions as a proxy for an Android smartphone. The authors do not explicitly motivate their decision for using a single-board computer instead of a real smart device. However, we assume that the Odroid-XU3's built-in power consumption monitoring tool played an important role. The number of papers utilizing a tablet is much less, only 4 of the 33 papers run their experiment on tablet. In addition there are 2 studies that employ both a tablet and smartphone giving us a total of 4 studies running their experiments on a tablet.

Table 6 only reports on the number of device types used not on the actual number of devices used. Some of the papers use more than one devices of the same device type to run their experiments on. In total (excluding the papers using both a tablet and a smartphone), 14

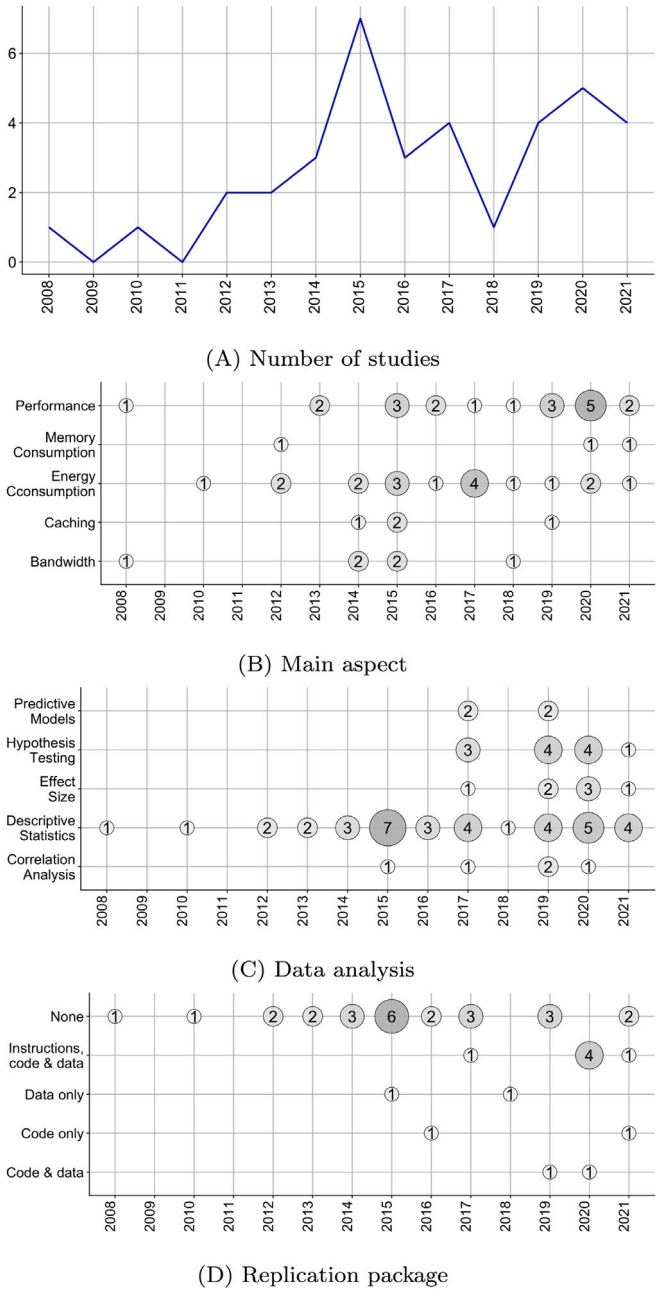


Fig. 2. Trend analysis (RQ1).

of the 33 studies use more than one device to run their experiments on. This is often done to get insights into the effect of different qualities of hardware on the measured results. For example, S13 runs their experiments on both a low budget and high budget flagship smartphone to see how this impacts the measured energy consumption and whether a difference between the two can be found. Similarly, in S5 4 different smartphones are used that were popular in 2015, 2016, 2017, and 2018, each running the most popular Android version in that year ranging from Android 4 to Android 7 which allowed the researchers to do historical studies.

Finally we can see that there are 6 papers that do not carry out their experiments on mobile devices at all, instead they use a form of emulation. Most recent browsers such as Google Chrome and FireFox are equipped with a set of developer tools that allow the user to simulate a range of other devices and browsers to approximate how

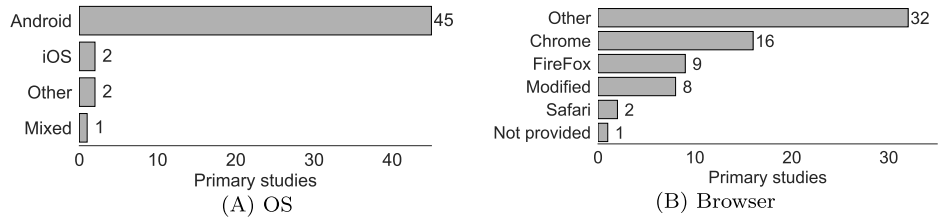


Fig. 3. Characteristics of the platform.

Table 7
Operating system.

Operating system	# Studies	Studies
Android	25	S1, S2, S3, S4, S5, S6, S7, S8, S10, S13, S14, S15, S18, S19, S20, S21, S23, S24, S25, S26, S28, S30, S31, S32, S33
iOS	1	S10
Other	1	S17
Mixed	1	S29

Table 8
Browser.

Browser	# Studies	Studies
Chrome	13	S1, S4, S5, S7, S8, S10, S13, S20, S21, S23, S31, S32, S33
Modified	8	S2, S3, S6, S15, S18, S25, S26, S29
FireFox	6	S1, S4, S20, S24, S28, S33
Safari	1	S10
Other	1	S17
Not provided	1	S19

the page looks and performs on a mobile device. For example, S15 uses a PC running the Chrome browser with its browser's emulation mode activated to make it act as an Android 4.2 native browse so that visited websites return their mobile-version of Web pages. The main reason given for using emulation is because it allows one to leverage the browser's programmability, which is not easy on real smartphones and tablets. Another interesting approach is taken by S8 as it makes use of the MONROE platform, which allows them to run measurements with full control of 100 nodes scattered in various locations across four different countries and connected via 11 commercial MBB providers. Each node consists of similar hardware to an average smartphone and is configured to mimic a mobile device browser by setting both the screen resolution and the user-agent accordingly. These studies all emulate a mobile browser. However, S28 emulates an entire operating system as it runs a Windows Phone 6 emulator on a desktop PC and runs a browser within that emulator.

Example. S32 focuses on low-end devices popular in developing countries that often suffer from poor web performance. To understand the root causes behind these suboptimal performances, they conduct an experiment to measure the memory utilization of popular websites across five different regions. They uncover that the primary culprit for hitting memory constraints is the execution of JavaScript code. Building on their results, they propose *WebMedic*, an approach that removes less critical functionalities of a webpage in exchange for improved memory utilization.

3.2.2. Operating system

Primary studies divided by employed operating system are listed in Table 7. The barplot in Fig. 3-A shows the frequency of operating systems used on the mobile devices (emulation based platforms were excluded) on which the measurement-based experiments were run. In total 50 mobile devices were used over the 33 primary studies. It can be observed that most devices use the Android operating system, namely 45 of the 50. Only 2 devices ran on iOS which were an Apple iPad 2 and an Apple iPhone 4 (S22). In the "other" category we found a device that was equipped with the Maemo 5 OS and one with the Symbian OS (S5). In one study (S29) end-users are involved in the experiment, conducted on their own device. Hence a "mixed" set of operating systems have been used in the experiment.

Example. The authors of S20 report that mobile Web page performance has improved over the years. However, it is not clear if these improvements are a result of better browsers, optimized Web pages, new platforms, or improved network conditions. To answer this question, they conducted a historical study over 4 years with 4 different operating systems (Android versions from 4 through 7) and multiple mobile browsers, measuring the effects of different factors on page load improvements. Their results highlight that the improvements in mobile page performance over the 4 years is largely due to improved platforms in newer mobile devices, and not a result of browser, network, or Web page improvements.

3.2.3. Browser

Studies listed by the browser employed during experimentation are provided in Table 8. Looking at Fig. 3-B we can observe that browsers falling in the category "Other" are the most used browser when doing measurement-based experiments on the mobile web. However, one study is responsible for the large number of browsers in this category, S5, as they tested 4 distinct versions of 6 different browsers falling in the other category. So, 24 of the 32 are the result of this study. Frequently used browsers in the "other" category are the native Android browser and Opera.

After that we see that Google Chrome is the most used browser (16 times) followed by Mozilla FireFox (9). In 8 of the 68 cases a modified browser was used. This was often done to make it easier to measure certain characteristics. For example, in S26 the authors added about 1200 lines of code to 27 files of a WebKit-based browser to make it possible to capture the dependency timeline. Apple's Safari is only used 2 times since only 2 of the 42 devices ran iOS.

Not shown in the plot but interesting nonetheless; for experiments using a form of emulation, Chrome is the most popular browser as its used in 3 of the 8 situations. Firefox is used 2 times and Opera just once

Example. *WebMythBusters* (S29) is a client–server application that users employ directly on their devices, to collect subjective metrics such as the user-perceived Page Loading Time (uPLT) and the Mean Opinion Score (MOS). The client has been developed using Kiwi browser, a modified version of Chrome that allows for extensions. The server governs the overall experiment, sending parameters to the client while receiving and storing measurement data.

3.2.4. Trend analysis

Below we report on the research trends over the years for parameters related to RQ2. Fig. 4-A reports the trend for the *device type* used in the experiments over the years. As can be observed, measurement-based experiments on the mobile web have been performed almost exclusively on smartphone devices, with 27 studies adopting this type of device. Only sporadically tablets or emulation-based devices have been used, with 4 and 8 usages respectively. This trend appears constant over the years.

Similarly, the trend for the employed *operating system*, visible in Fig. 4-B, reports an almost exclusive usage of Android (48 usages), with occasional usage of iOS (2 usages). Two usages of other operating systems have been observed only in the year 2010 when the current market share dominance of Android and iOS was less marked.

A diverse trend can be observed for the *browser* parameter, displayed in Fig. 4-C. Usage of Chrome has been ongoing since the year 2013 and has been experiencing a growing trend. Similarly, FireFox has been experiencing an increasing experimental adoption in the more recent years, with 5 papers reporting its usage in the 2019–2021 period. An opposite trend can be observed for modified browsers, that while consistently used until the year 2017 has since experienced a decrease, with only one study reporting the usage of a modified browser in the year 2021. This trend is potentially due to increased customizability and extensibility of more recent Chrome and FireFox releases.

Summary of the main findings (RQ2):

- Smartphones are the most used types of devices, followed by emulated devices; tablets are seldom used.
- Android is the leading operating system in measurement-based experiments on the mobile Web. Only one recent study covers both Android and iOS.
- The used browsers follow the Android market share, where Google Chrome is the most used browser, followed by Mozilla Firefox. Other types of browsers (*i.e.,* Apple Safari) are seldom used.

3.3. Subjects (RQ3)

3.3.1. Type

Table 9 shows the frequency of the types of website considered across the 33 primary studies. It can be observed that 19 of the 33 studies use real websites, i.e. no toy examples, no demo Web apps, no Web apps developed by students or non-professional developers and no Web apps specifically created for the experiment. However, 14 studies make use of synthetic websites. Studies using both real and synthetic websites often created synthetic copies to see how certain changes on real Web pages impacted the measurement results. For example, S24 first measures the energy consumption of 25 top websites. After that they look at the energy consumption of individual Web elements by copying the Web pages and commenting out specific components to see how this impacts the energy usage. The 5 papers that exclusively used synthetic Web apps created their own Web apps specifically for the experiment. For example, S9 built a Web app that used three different web-based communication protocols (polling, long polling and websockets) to see how they compare in terms of energy consumption. Similarly, S27 measured the energy consumption of a simple Web app that was implemented in 8 different JavaScript frameworks and programmed by computer science students at the master and post-graduate level.

Example. S1 analyzes how device memory usage affects Web browsing performance, by measuring the memory footprint of the top 100 Web pages from Alexa over different mobile browsers. Afterwards, a deeper analysis of memory usage in Chrome is conducted employing a set of specifically crafted synthetic websites, each built to exercise a specific part of the web browser. By building on the collected evidence, the authors propose a set of optimizations that can improve performance and reduce the chances of browser crashes in low-memory scenarios.

Table 9
Subjects type.

Type	# Studies	Studies
Real	19	S2, S3, S5, S6, S7, S9, S10, S12, S16, S18, S20, S21, S22, S24, S26, S29, S30, S31, S32
Both	9	S1, S8, S11, S13, S14, S15, S19, S23, S25
Synthetic	5	S4, S17, S27, S28, S33

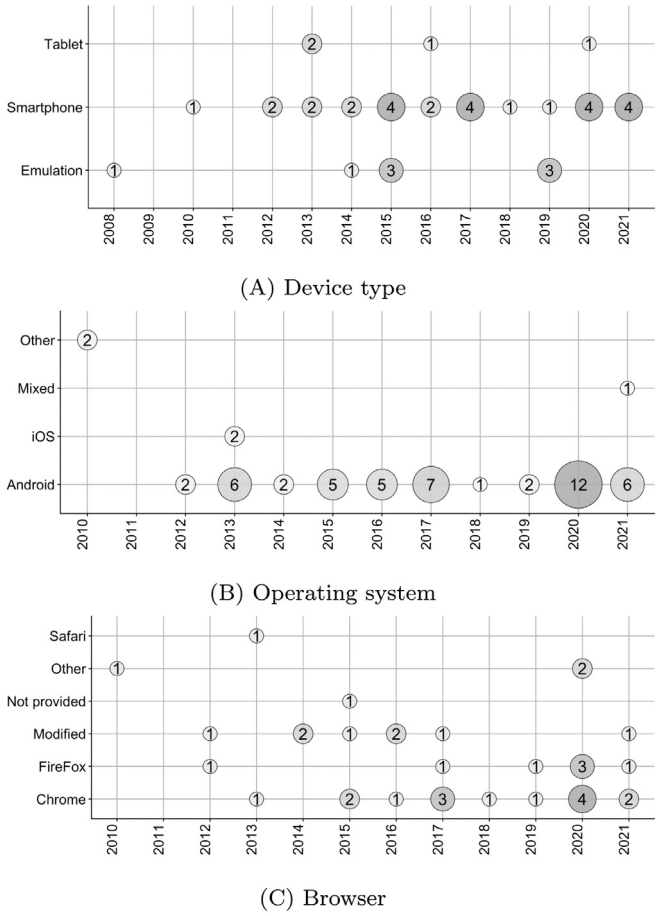


Fig. 4. Trend analysis (RQ2).

3.3.2. Selection

The plot shown in Fig. 5-A shows how the real websites were chosen. Some papers use more than one source, so the number of occurrences does not correspond to the number of primary studies, that are shown in Table 10. Noticeable is the high prevalence of Alexa as a source to select websites, 15 of the 33. In 8 cases it was not explicitly mentioned where the website selection is based upon. For example, S19 uses both the New York Times and the Web page of their university but no motivation is given. Four others used another source, specifically: S26 based their website selection on a blog post listing the 10 most visited websites on mobile phone in 2009. S13 and S4 both selected Web apps from a repository of PWAs called PWARocks. In

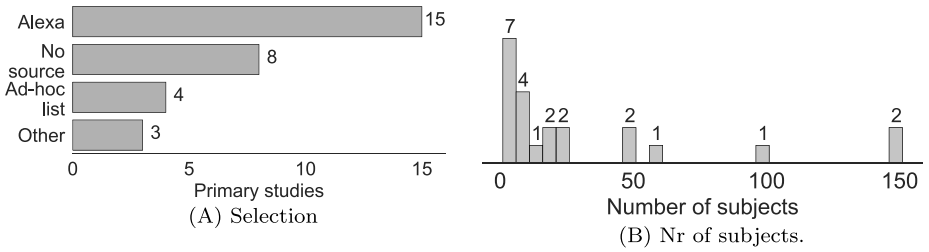


Fig. 5. Characteristics of considered subjects.

Table 10
Subjects selection.

Source	# Studies	Studies
Alexa	15	S1, S2, S3, S5, S8, S9, S11, S12, S13, S18, S20, S21, S22, S26, S32
No source	5	S3, S10, S14, S15, S19, S25, S29, S31
Ad-hoc list	4	S6, S23, S24, S30
Other	3	S7, S8, S16

Table 11
Subjects hosting.

Hosting type	# Studies	Studies
Original	19	S2, S3, S5, S6, S7, S8, S9, S10, S11, S12, S14, S16, S18, S20, S21, S22, S29, S30, S31
Mirrored	8	S1, S13, S15, S23, S24, S25, S26, S32
Not provided	1	S19

Table 12
Number of subjects.

# Subjects	# Studies	Studies
# ≤ 10	11	S2, S4, S7, S10, S14, S15, S17, S23, S24, S27, S28
10 < # ≤ 50	7	S5, S6, S8, S13, S19, S22, S25
50 < # ≤ 100	2	S1, S29
100 < # ≤ 1000	7	S9, S12, S18, S20, S21, S26, S30
# > 1000	2	S11, S16
Not provided	4	S3, S31, S32, S33

the `other` category we found one paper that used a set of Javascript benchmarks (S22), a paper for which the website was provided prior to the experiment (S3), and one that scraped Web pages meeting certain requirements (S7).

Example. S16 presents a characterization of Google's Accelerated Mobile Project (AMP) impact on users' QoE. The authors compare a corpus of over 2,100 AMP webpages and their corresponding non-AMP counter-parts, scraped starting from an initial list of 578 keywords found using Google Trends. Their results show that AMP significantly improves the webpages SpeedIndex at the cost, however, of an average 1.4 MB of additional data downloaded, unbeknownst to users.

3.3.3. Hosting

Table 11 depicts the way real websites (so excluding the two papers that only use synthetic websites) are hosted. It can be observed that most papers prefer to use the original websites' host. However, in 8 of the 28 cases the website(s) were mirrored. This is often done as the researchers want to create a fully controlled environment. They might for example simulate certain network conditions, hosting a website on a server that is in their control makes this easier. One study did not explicitly report how the websites were hosted (S19).

3.3.4. Number of subjects

The frequency of the number of subjects used in an experiment is shown in Fig. 5-B using a histogram with the bin width set to 5 and summarized in Table 12. For 4 studies it was not clear how many subjects were used in total.

We can see that 22 of the 33 studies use 150 subjects or less in their experiment. 7 of these 22 papers use 5 subjects or less. The 11 remaining papers are not included in the graph for the sake of readability as the number of subjects used differs significantly. For example, there are two studies that use 3400 and 95,728 subjects respectively (S7, S18). This large number is primarily caused by the fact that analysis was done afterwards. For example, S18 examines 95,728 websites by crawling all these websites and downloading all assets loaded by each website, along with a record of all request and response details saved in an HTTP archive record (HAR) file. Then later, all this data is analyzed. Four studies (S3, S31, S32, S33) do not report the number of subjects involved in their experimentation.

3.3.5. Subjects provided

In total 15 of the 28 papers that used real websites provide the actual URLs to these websites. This is often done by means of an appendix, a file in the replication package or an in-text table listing all the websites.

3.3.6. Trend analysis

Focusing on the *subjects type* used in mobile web experiments, observing the yearly grouping of studies provided in Fig. 6-A, we can observe a constant trend regarding the usage of real subjects over the years. However, we can also observe a minor reduction in the number of studies that employed synthetic subjects, with only 3 papers in the past four years employing them.

Regarding the *subjects selection*, depicted in Fig. 6-B, the trend over the considered years highlights a reduction in the number of studies that do not report the source of their subjects, in favor of more studies that report the Alexa list as the source of their subjects. Overall we consider this as a positive trend, as reporting the source of the subjects improves the replicability of the studies.

In regards to the subjects *hosting*, pictured in Fig. 6-C, a constant trend can be observed, with studies mostly employing subjects on their original hosting. This trend negatively impacts the replicability of the studies, as websites frequently change over time, as well as limiting the amount of control exercisable over the experimental environment, as previously discussed in Section 3.3.3.

The yearly breakdown for the *number of subjects* used in experiments is plotted in Fig. 6-D. We can observe a reduction, albeit moderate, of the studies that employ a large number of experimental subjects. Indeed, in the 2019–2021 period, only 4 experiments have used more than 100 subjects.

Finally, from Fig. 6-E, a constant trend can be observed for the *subjects provided* parameter, with a considerable number of studies

Table 13
Experiment scope.

Scope	# Studies	Studies
Page load	25	S2, S4, S5, S6, S7, S8, S9, S10, S12, S13, S15, S16, S18, S20, S21, S22, S24, S25, S26, S27, S28, S29, S30, S31, S32
Usage scenario	5	S11, S14, S17, S23, S33
Both	3	S1, S3, S19

that do not provide the actual URLs to the websites used during the experimentation.

Summary of the main findings (RQ3):

- Researchers tend to use real subjects in their experiments, and in fewer cases they use also synthetic subjects.
- In the past, researchers tended to do not report the sources from which their subjects were sampled. Recently, despite its known limitations [27], the Alexa list is the most used source for real subjects.
- Subjects are generally hosted on their original servers, which benefits the external validity of the experiments (possibly at the expense of its internal validity).
- Experiments frequently have a low number of subjects (*i.e.,* less than 10), but there is also a good number of studies having a high number of subjects (*i.e.,* more than 50).
- Despite the clear advantages in terms of experiment replicability, there is still a certain balance between studies explicitly reporting the considered subjects and those not reporting it.

3.4. Experiment execution (RQ4)

3.4.1. Scope

Table 13 summarize the scope of the experiments. Most studies focus on page load only, namely 25 of the 33. In total 8 papers actually experiment with usage scenarios. For instance S18 simulated a few user interactions to invoke additional content and functionality that initially may be hidden after loading the page to measure bandwidth usage. Three of the 8 papers do both experiments focusing on page load and usage scenarios. In S1 the authors also provide insights into the effect on memory usage when a user scrolls a page and uses multiple tabs in addition to just loading the page.

Example. S19 discusses how to refine mobile web design for reducing the energy consumption of web browsing on mobile terminals. By means of experiments, the authors analyze the energy consumption of different computing resources on mobile terminals during web page browsing. Their results reveal scrolling operations play an important factor in energy consumption which, has not been considered in previous works. Based on the fact that web contents have different access popularity, they propose a content rearrangement method: listing contents in the decreasing order of their access popularity to reduce the average number of scrolling operations required to reach target contents.

3.4.2. Focus

Table 14 lists primary studies by their experimental focus. Out of the 33 primary studies, 28 focus on Web pages as a whole, they do not put a special emphasis on one of the three essential Web technologies (HTML, CSS, JavaScript). The 5 exceptions all focus on JavaScript (S9, S27, S13, S24, S32). For example, S27 implemented a simple Web app in 8 different JavaScript frameworks to see how each framework influences the energy consumption.

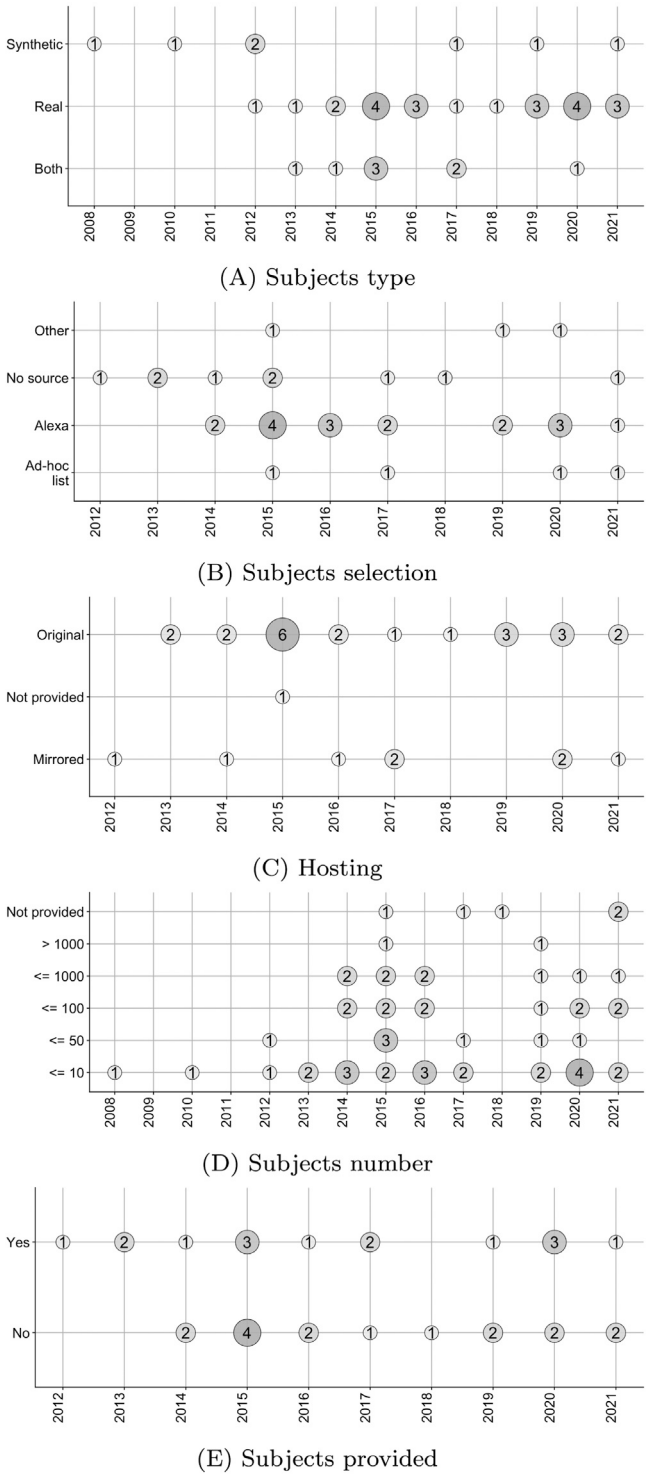


Fig. 6. Trend analysis (RQ3).

Example. S17 investigates the battery consumption of JavaScript applications running on mobile phones. In their empirical study, eight implementations of the same application – each using a different JavaScript library – were developed and analyzed. The results highlight

Table 14
Experiment focus.

Focus	# Studies	Studies
Whole page	28	S1, S2, S3, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S18, S19, S20, S21, S22, S24, S25, S26, S27, S29, S30, S31, S33
JavaScript code	5	S4, S17, S23, S28, S32

that there are significant differences between different implementations and no single factor is enough to explain the performance differences.

3.4.3. Tools

The used instrumentation to do the measurements is of course dependent on what is exactly measured. For the total of 16 papers that measure energy consumption 10 do that by using software based tools. For example, S9 uses two different software based energy profilers: the Trepn profiler and the GreenSpector profiler. The study states that while hardware based profilers usually offer higher precision, selecting and configuring hardware equipment is complex and can therefore introduce additional bias. Additionally, it requires special equipment which makes reproduction of the experiment more difficult. Another interesting example, S21, feeds tcpdump traces into a radio energy model to get the energy consumption. The other 4 studies use hardware based measurement tools such as the Monsoon power monitor or other multimeters. One paper (S11) argues that while using software based tools results in a simpler and cheaper measurement setting they have several drawbacks. For example, the energy software may be available only for certain mobile devices or Operating Systems, they can cause a form of energy overhead and thus biasing the measurement results or the accuracy of the results strictly depends on the supported power models and the implemented APIs.

For the 19 papers that measure performance its more difficult to pinpoint overarching themes as the tools used are very diverse. When mirroring the original pages some studies inject custom written JavaScript code to measure the PLT (S8, S4, S30). Similarly, S5 uses Boomerang, a JavaScript library that measures performance timings, metrics and characteristics of your user's Web browsing experience,[6] when its embedded into the page. Noticeable is the high prevalence of tools created by Google. Two papers (S17, S1) use Telemetry[7] a performance testing framework that allows users to perform arbitrary actions on a set of Web pages (or any android application) and report metrics about it. Two other papers (S3, S2) use Google Lighthouse.[8]

To measure bandwidth, tools to inspect network traffic like WireShark are often utilized. Of the three papers that measure memory consumption one created their own app to measure PPS (S6), a second one used Google Chrome's Timeline Tool (S25), while the third relied on the dumpsys Android utility (S32).

Finally we see a lot of custom created tools that manage the orchestrating process. For example, the authors of S26 created an Android application that opens the system's default browser and visit a pre-provided list of websites. Other tools frequently used are proxies like Charles and Fiddler and Linux Traffic Control (tc) to simulate network conditions.

Example. S6 focuses on the internals of web browsers on smartphones, using the WebKit codebase, two generations of Android smartphones, and webpages visited by 25 smartphone users over three months. To do so the authors implemented a smartphone tool called *PageCycler* to visit URLs via the smartphone browser while recording the network traffic. Their results demonstrate how the internals of browsers and operating systems contribute to the page load delay and therefore reveal opportunities for optimization.

[6] https://github.com/akamai/boomerang

[7] https://chromium.googlesource.com/catapult/+/HEAD/telemetry/README.md

[8] https://developers.google.com/web/tools/lighthouse

Table 15
Network condition.

Network	# Studies	Studies
WiFi	17	S3, S4, S5, S13, S14, S15, S17, S18, S19, S22, S23, S24, S26, S28, S30, S31, S32
3G	9	S3, S6, S14, S15, S17, S18, S22, S25, S26
Not provided	5	S1, S2, S8, S9, S10, S11, S29
4G	4	S7, S14, S22, S26
Ethernet	4	S6, S12, S21, S22
Simulated	2	S16, S20
LTE	1	S13
2G	1	S23
GPRS	1	S27
Not applicable	1	S33

Table 16
Caching.

Caching status	# Studies	Studies
Disabled	13	S1, S3, S5, S7, S8, S13, S14, S15, S20, S22, S23, S28, S32
Not provided	11	S2, S4, S6, S11, S16, S17, S19, S26, S27, S29, S33
Both	8	S9, S12, S18, S21, S24, S25, S30, S31
Enabled	1	S10

3.4.4. Network conditions

Fig. 7 shows the network conditions under which the experiments took place. Again, since some studies experimented with multiple network conditions the number of occurrences does not correspond to the number of primary studies, that are listed in Table 15.

It can be observed that most experiments use a real, non-simulated, WiFi network (15 occurrences) followed by a real 3G network (8 occurrences). The real Ethernet connections are primarily because of studies that used a form of emulation, i.e., using a Desktop and mimicking a mobile browser.

When simulating network conditions, often (5 times) the exact network type is not defined, instead only the down- and upload speeds are given. In 7 cases the paper did not elaborate on the network conditions under which the experiment was carried out. One study (S33) is excluded from this categorization as no network requests are fired during their experimentation.

3.4.5. Caching

We can observe from Table 16 that most, 13 of the 28, primary studies disabled caching during their experiments to make sure that all requested data will be from the server. This makes sure that for each run of the experiment the same environmental conditions hold. One exception is S22 where the authors conclude that the loading time required for ten and twenty images is similar after the second run because they did not clear the cache.

Eight primary studies did experiments with both an enabled and disabled cache. Most of these experiments do this to assess how caching influences the characteristic under measurement. For example, S21 does two types of loads. A cold-cache load where all caches are cleared

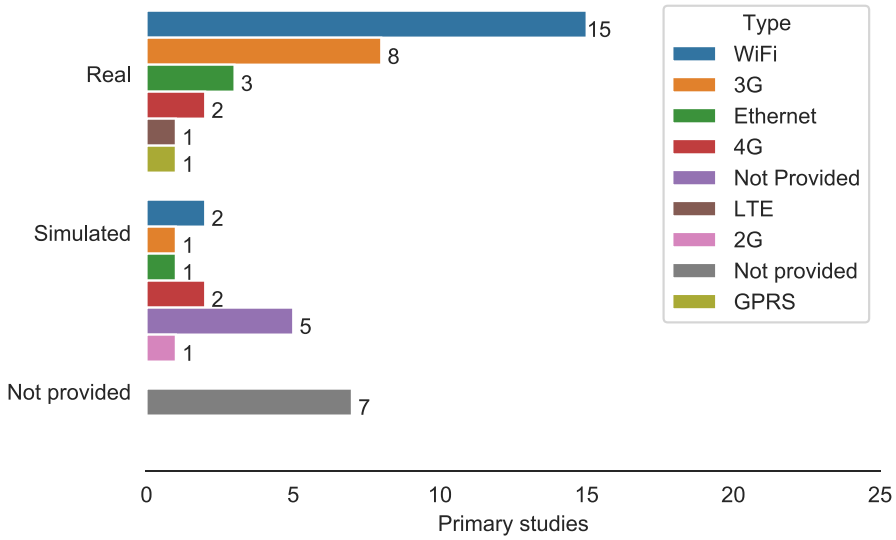


Fig. 7. Network conditions while running the experiment.

before loading, and a warm-cache load right after the cold-cache load without clearing any cache. Finally we can see that 11 studies did not give any information about whether their experiments ran with cache enabled or disabled.

Example. S24 aims at assessing the impact of caching on both the energy consumption and performance of Progressive Web Applications (PWAs). To do so, the authors conducted an empirical experiment targeting 9 real PWAs developed by third-party developers. The experiment is designed as a 1 factor - 2 treatments study, with the usage of caching as the single factor and the status of the cache as treatments (empty vs populated cache). Their results show that PWAs do not consume significantly different amounts of energy when loaded either with an empty or populated cache. However, the page load time of PWAs is significantly lower when the cache is already populated.

3.4.6. Trend analysis

With respect to the *experiment scope*, we report the yearly groupings in Fig. 8-A. From the collected data emerges a constant trend across the years: *Pageload* is the scope researchers have been mostly focusing on, while usage scenarios, to investigate actions triggered after the page has loaded, are seldomly used across all years. Similarly, a constant trend can be observed for the *experiment focus* parameter, displayed in Fig. 8-B: most studies have focused on the web page in its entirety, and only a minority has investigated exclusively the JavaScript code embedded in the page.

Focusing on the *network status* parameter, plotted in Fig. 8-C, we can observe that several different connection technologies have been used over the years. In particular, we can observe a reduction in the studies that use older connection technologies, in favor of newer ones. A glaring example is a 3G technology, with no studies using it since the year 2019, likely supplanted by the more recent 4G standard, which indeed has started to be used in studies from the same year. However, slower connections (such as LTE or 2G) have been used in the year 2017, to purposely conduct experiments in conditions of limited available bandwidth. The majority of the studies however relied on WiFi connections, a trend that holds in more recent years.

Finally, focusing on the *caching* parameter, we provide the yearly data in Fig. 8-D. From it, we can observe in recent years an increase in papers that have disabled the cache while performing experiments, alongside a reduction in the number of papers that have experimented with both caching enabled and disabled (6 papers in the 2014–2016 period, only 3 papers in the 2019–2021 period). This trend highlights that researchers are increasingly aware of the impact that caching can have on measurements performed during experimentation.

Summary of the main findings (RQ4):

- The vast majority of the experiments are performed considering only the page load of the measured Web apps, leaving out the overall experience of the users navigating through them.
- The vast majority of the experiments focus on the whole page (HTML, JSS, CSS), while a minority focusses on the JavaScript code.
- WiFi is the most used network condition in the analyzed studies, followed by 3G and 4G. A non-negligible number of studies do not report the network conditions under which the experiment has been carried out.
- Caching is either disabled completely or it is both enabled and disabled, depending on the factors and treatments of the experiment. A non-negligible number of studies do not report whether the cache is cleared/disabled/considered across the various runs of the experiment.

4. Horizontal analysis

This section reports on the results of our horizontal analysis. It is worth recalling that, in this phase of the study, we (i) built contingency tables for pairs of parameters coming from our vertical analysis, (ii) analyzed each one of them, and (iii) identified perspectives of interest.

4.1. Main aspect — Subjects selection

Fig. 9 plots the analyzed studies by their main focus opposed to the kind of subjects used in the experimentation. From it, we can observe an overall balance across considered aspects, with memory, energy, and networking that have a close to equal proportion between studies that have used real subjects and studies that have used synthetic ones. We speculate that, since measurements for these aspects are harder to collect, researchers feel more the need of crafting their own experimental subjects in order to conduct measurements more easily. However, performance-oriented studies have a strong prevalence of real web apps (18/33 studies) with respect to synthetic web apps (5/33

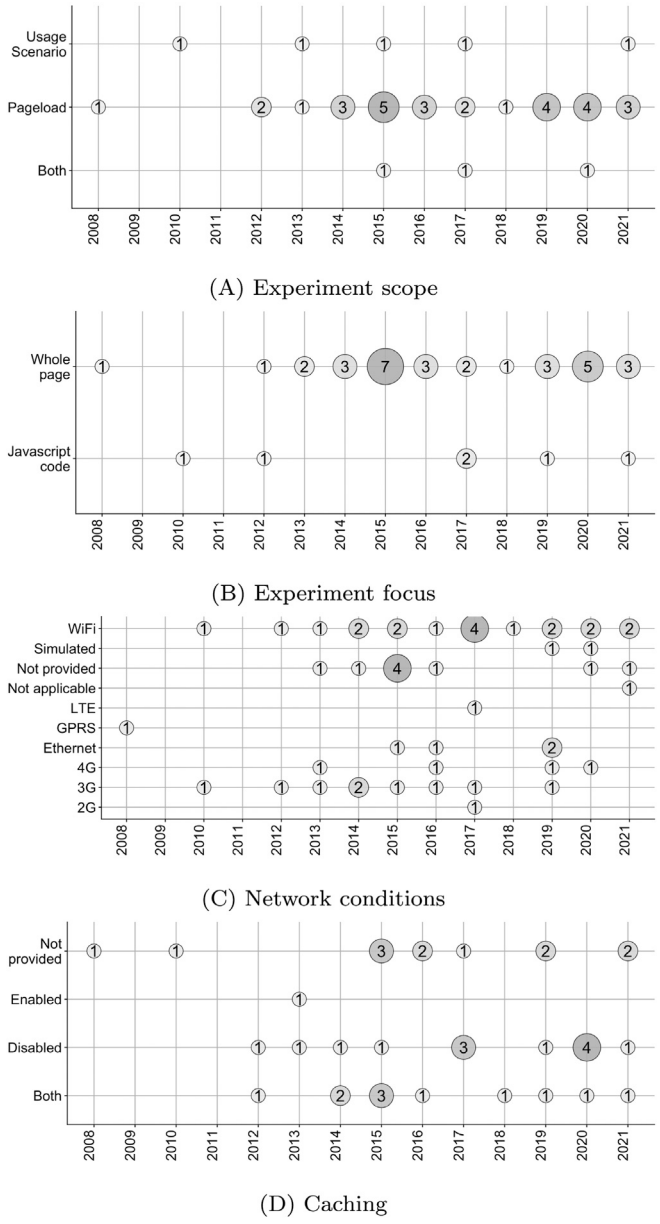


Fig. 8. Trend analysis (RQ4).

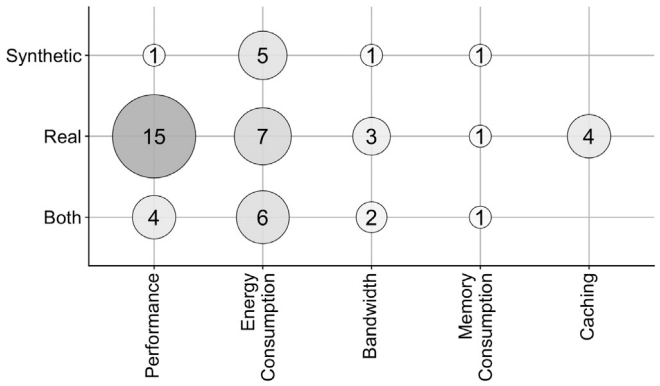


Fig. 9. Main aspect — Subjects selection.

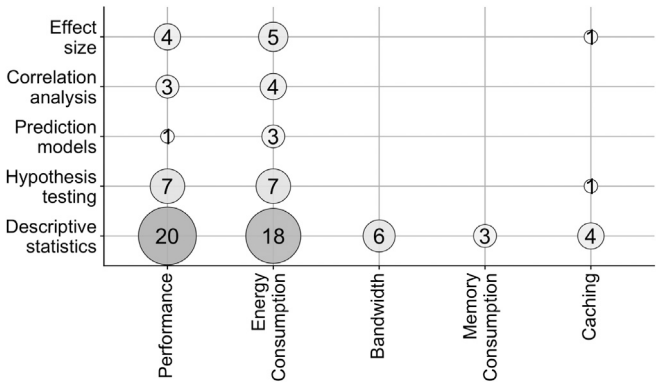


Fig. 10. Main aspect — Data analysis.

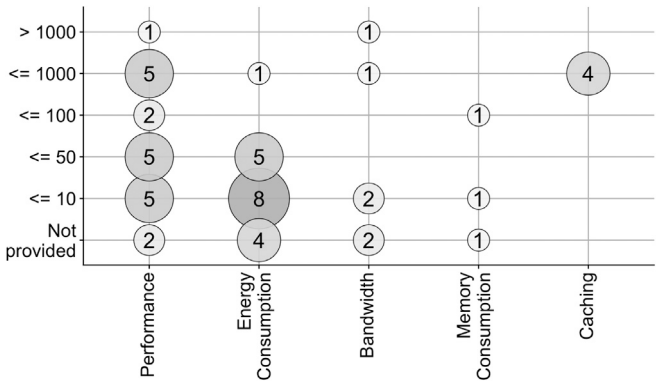


Fig. 11. Main aspect — Subjects number.

studies). Better understanding of why this phenomenon is happening is interesting. Noticeably, the studies that focused on caching have used exclusively real subjects in their experimentation.

4.2. Main aspect — Data analysis

A visualization of studies divided by considered main aspect and by used analysis techniques is given in Fig. 10. As expected, descriptive statistics and hypothesis testing are the most used data analysis strategies across all studies main aspects. However, all studies on memory consumption and networking use descriptive statistics only (no effect size estimation, no hypothesis testing, no correlation analysis and no use of prediction models). We do not have a clear explanation for this phenomenon though.

4.3. Main aspect — Number of subjects

An overview of studies grouped by aspect and number of subjects is provided in Fig. 11. Interestingly, measurement-based experiments on energy consumption tend to involve a lower number of subjects than the other types of experiments (*e.g.,* performance). Indeed, the majority of experiments focusing on energy have less than 10 subjects (8 cases over 19), followed by having less than 50 subjects (5 cases over 19). This result might be seen as an indication of the effort and time required to execute energy-related experiments, which are notoriously more demanding than other types of experiments [28]. One primary study (S18) on energy is considering a high number of subjects, likely due to the fact that the study relies on network traces to estimate the energy consumption caused by network transfers (and thus can be more easily scaled). Nonetheless, the authors developed more than 4 thousands lines of code to fully automate the experiment execution.

Similar considerations can be done for memory consumption, for which two out of the three studies focusing on this aspect have employed less than 100 subjects, while the third one does not report the

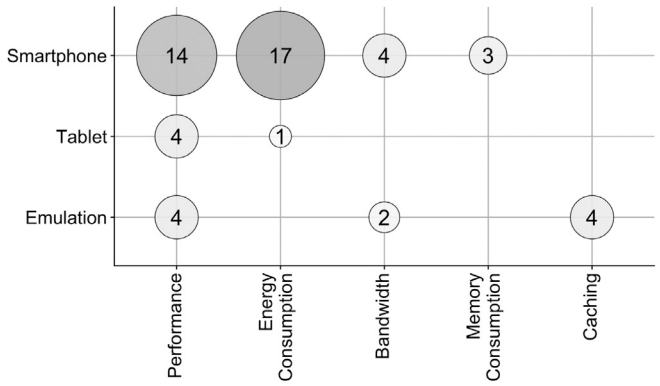


Fig. 12. Main aspect — Device type.

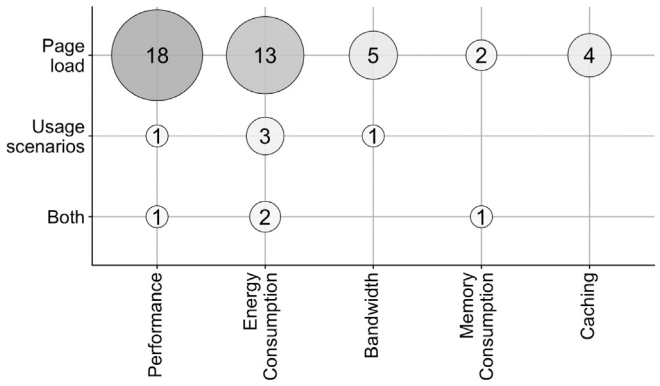


Fig. 13. Main aspect — Experiment scope.

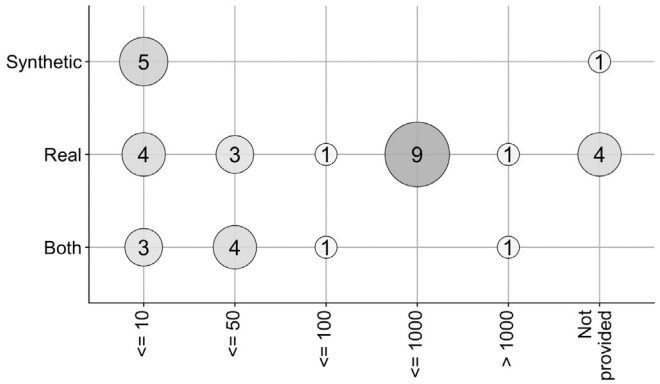


Fig. 14. Subjects number — Subjects type.

number of subjects involved in the experiments. On the other side of the spectrum, all studies that focus on caching employ a higher number of subjects, with over 100 subjects for all the four studies. We hypothesize that the reason for the high number of subjects in these studies is due to the simpler effort required to conduct measurements. Indeed, caching-related metrics are easier to collect, as cache hits and misses can be obtained from HTTP traffic logs. Moreover, a minor effort is required to restore the experimental environment to initial conditions between runs (clearing the browser cache is sufficient), contributing to the greater scalability of these experiments.

4.4. Main aspect — Device type

Studies by main considered aspect and by device type are presented in Fig. 12. It stands out that no energy-oriented studies are using emulation, which is understandable as performing energy measurements on an emulated device can be imprecise, leading to unreliable results. However, we can observe the presence of performance-oriented studies that have been performed on emulated devices. Similarly to energy consumption, it is well-known that emulators are not meant to have a representative performance of the emulated devices, and thus it can severely affect the collected results. However, none of the involved studies (S12, S16, S22, S27) report this fact as a potential threat to validity.

4.5. Main aspect — Scope

Studies grouped by their main aspect and by their experimental scope are provided in Fig. 13. From the plot, it can be noticed that usage scenarios have been more commonly used in energy-oriented studies (3/33) rather than in studies focusing on other aspects (1/33 for networking and 1/33 for performance).

This result might be explained by the fact that energy consumption is strictly dependent on the total amount of time an operation is performed ($E = P * t$) and that in order to build a proper assessment of the energy consumed by a Web app researchers cannot always rely on a few samples purely found *"in the wild"*, but it is often necessary to construct ad-hoc scenarios. Indeed, all the three involved studies (S17, S23, S33) investigate the differences in energy consumption of different development approaches and such comparison would be difficult without crafting tailored experimental subjects.

4.6. Number of subjects — Subjects type

Primary studies by number and type of subjects are provided in Fig. 14. It does not come as a surprise that experiments with synthetic subjects tend to involve a fewer number of subjects, with 8 cases with less than 10 subjects, 4 having less than 50 subjects, and only 1 experiment having more than 100 studies.

The fact that synthetic subjects are used in experiments with fewer subjects might be an indication of the fact that developing synthetic subjects is time consuming for researchers, who cannot afford to invest time to developing hundreds of synthetic subjects for their experiments. As a matter of fact, in the only study with more than 100 synthetic studies (S11) a number of compression tools are used on real website to obtain the synthetic ones, in order to see how the bandwidth use is affected. Hence, only a simple transformation is applied to craft the synthetic websites. Therefore, going forward, the automatic generation of more complex synthetic subjects might be a promising research direction to be investigated.

4.7. Number of subjects — Device type

A breakdown of studies by number of subjects and by device type is provided in Fig. 15. It can be observed that emulation-based experiments tend to be used more in experiments with a higher number of subjects. Indeed, differently from experiments with real devices, where we see a prevalence of experiments with less than 10 subjects, experiments with synthetic subjects tend to be more used when more than 50 subjects are considered: out of the 8 studies using emulation, 4 have more than 100 subjects, 2 more than 1000 and only 2 having less than fifty and less than 10 respectively.

This phenomenon might be an indication that emulation-based experiments, which generally last longer, can scale in an easier manner to a higher amount of subjects. Also, among all considered primary studies, there is not a single experiment involving a real device and more than 1000 subjects; this is a confirmation of the intuitive perception that experiments performed on real devices do not scale. In some cases,

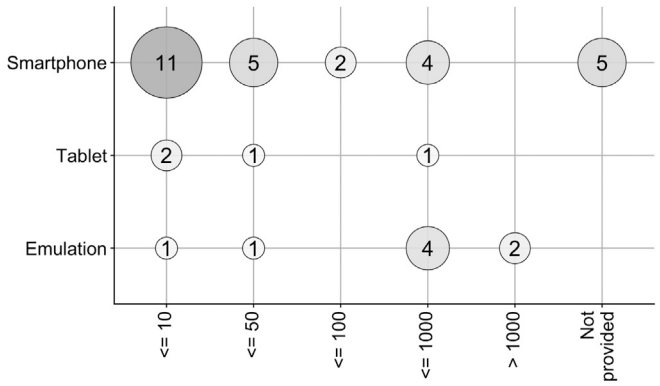


Fig. 15. Subjects number — Device type.

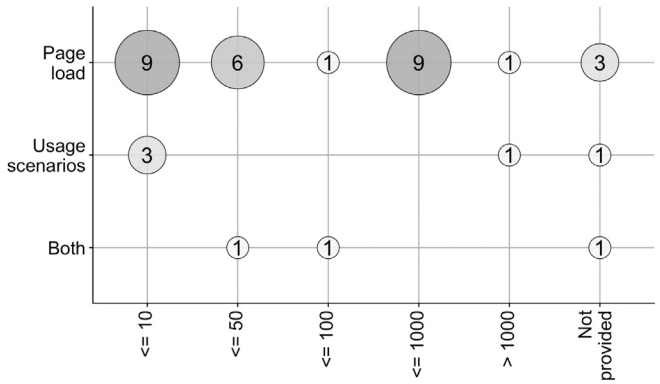


Fig. 16. Subjects number — Experiment scope.

this limitation can impact the validity of the study, mostly in terms of low statistical power and external validity of the results. Researchers tend to mitigate the potential bias of having a low statistical power by repeating the measures for each trial of their experiment, whereas the bias with respect to the external validity of the experiment is accepted and reported in the discussion of the threats to validity of the considered studies.

4.8. Number of subjects — Experiment scope

Fig. 16 reports on studies grouped by number of experimental subjects and experimental scope. Out of six experiments involving the execution of usage scenarios and reporting the number of subjects, three (S14, S17, S23) have less than 10 subjects, one (S19) has less than 50 subjects, one (S1) has less than 100 subjects, and one (S11) has more than 1000 subjects.

This distribution of involved subjects suggests that the design and execution of usage scenarios tend to take longer than experiments focusing on page load. Indeed, for all the studies in which more than 10 subjects are used only simple scenarios are executed. In fact, S11 employs more than 1,000 subjects during experimentation but the used usage scenarios only perform basic actions (page scrolling, mouse events like click and hover) to invoke additional content and functionality after completing page load. Similarly, the other two studies that employed usage scenarios and a considerable number of subjects (S1 and S19) only perform some page scrolling after page loading has been completed. In contrast, studies that employ a more limited number of subjects employ more complex scenarios, that have been specifically crafted for each experimental subject. For instance, S14 reports to "*have generated 4 trace files, each targeted at a particular website. The Amazon trace browses products from the amazon.com US store. The Craigslist trace searches for various items in the Western Massachusetts Craigslist website. […]*". Hence, this confirms the intuition that designing complex execution scenarios is currently a challenge for researchers when a large number of subjects is involved in the experimentation.

5. Discussion

In this section we discuss the main insights emerging from the results of each research question and make recommendations for both researchers and Web developers.

5.1. Insights and recommendations about metrics and data management (RQ1)

Regarding the distribution of considered aspects in the primary studies, we can say that **the focus of research lies primarily on measuring the performance and energy consumption**. This result is not surprising as these two characteristics are very noticeable by users and thus have a great influence on the QoE which in turn, as previously discussed, is vital towards the success of mobile Web apps [29,30]. However, as a consequence, a relative limited number of measurement-based experiments have been performed on caching, memory consumption, and bandwidth usage. We invite researchers to investigate on those aspects as well since they also contribute towards the perceived QoE of mobile websites. Therefore, **more research into cache performance, memory consumption, and bandwidth usage on the mobile Web is advised**.

Below we report our recommendations about the used metrics. We suggest to **use Joules for energy consumption** since it is used in the majority of analyzed primary studies and is widely accepted and understood within the software engineering community. In our set of primary studies we have a relative limited number of papers measuring bandwidth, memory consumption, and caching, therefore our recommended metrics for these characteristics can be less reliable. Nonetheless, we recommend to measure these characteristics using bytes, including derivatives like Proportional Set Size (PSS) and hit rate, respectively.

The landscape of available metrics for performance is highly fragmented. We presume that this is the result of the introduction of the JavaScript Performance APIs, such as the Performance Timeline[9] and Navigation Timing API,[10] in combination with the development of tools like Google Lighthouse, WebpageTest and SiteSpeed.io that allow developers to assess the performance of their Web apps via their own metrics. Page Load Time is still a solid metric for the majority of use cases. However, some papers argue that other metrics such as SpeedIndex and above-the-fold time represent user-perceived load times better [12,31]. While the choice of a metric depends strongly on the context and goal of the experiment, we advice researchers and practitioners to **clearly define and explain the chosen metrics and, if possible, to explicitly describe how it differs from other popular metrics as well**.

Almost two third of the primary studies report their results without carrying out hypothesis testing. This result is not worrisome *per se* since (i) studies can have an exploratory nature and (ii) the blind application of statistical tests might be problematic as well [32,33]; however, several primary studies draw strong conclusions and even make recommendations based on their collected data, without providing evidence about whether it is statistically significant or not. Adding to that, only 6 papers do effect size estimation which is often considered to be essential when reporting the results of a statistical analysis [34]. We recommend researchers to **carry out and report in details a proper statistical analysis of the obtained measures, when the goal of the**

[9] https://www.w3.org/TR/performance-timeline
[10] https://www.w3.org/TR/navigation-timing

study aims at establishing evidence about a given phenomenon. The low amount of papers utilizing correlation analysis and predictive modeling is less worrisome as the they are often only applicable in certain contexts.

Alarming is the low number of studies providing a replication package. This is unfortunate as replicability is considered to be the major principles of the scientific method and its importance has been emphasized multiple times over the years [35–37]. Ideally, scientific results are documented in such a way that their independent verification and replication is fully possible. It may be interesting to look into defining a set of guidelines for replication packages provided by studies doing measurement-based experiments in the context of the empirical software engineering field. On a broader scale there are several existing initiatives like the Open Science movement with its Open Science Framework and rOpenSci which provides a reproducibility guide, including a checklist from which inspiration can be drawn[11] [38].

5.2. Insights and recommendations about the platform (RQ2)

The distribution of device types shows a relative low number of experiments run on tablets. This may be a problem since we can argue whether the conclusions drawn based on smartphone based experiments carries over to tablets. In this regard it would be advised to stimulate the use of tablets as experimental environments, depending on the scope of the experiment being carried out. Nevertheless, as discussed in the introduction of this paper, 55% of all worldwide Web traffic came from mobile devices. However, only 2.83% of that consisted of traffic from tablets, the other 52.95% is from smartphones [1]. This can explain and justify the high number of experiments using smartphones.

We applaud the relative high number of studies using more than one device in their experiments; **using more than one device improves the generalizability of the results of the experiment**. While it is generally preferred to use real devices to perform experiments, we understand that the use of emulation can be advantageous in some situations. One of the primary studies mentioned the reason for using a Desktop browser in emulation mode is to make it easier to use the programmability of the browser. However, over the years lot of new tools have been developed to which the required programmability can be possibly delegated as they help streamlining the orchestrating process of setting up and executing measurement-based experiments. A good example of such tools is Android Runner, an extensible framework for automatically executing measurement-based experiments on native and Web apps running on Android devices [39]. Finally we have found that most papers do not explicitly motivate their choice for a certain device or combination of devices. Although we understand that resources are often scarce making the device type a given, we feel that providing a rationale could lead to more insights.

Android is clearly the most used operating system when doing measurement-based experiments on the mobile Web. This corresponds to the latest trends concerning the worldwide mobile operating system market share as of November 2020 where Android is responsible for 71.18% of the market [40]. Another reason that may have contributed to the high prevalence of the Android operating system is that it is open source and provides a wide software ecosystem. This enables researchers and practitioners to use it as an environment for experiments since it imposes no restrictions on the applications the user can install. This in contrast to for example iOS. However, this does not completely justify the low number of studies doing experiments on devices running iOS. iOS is still covering 28.19% of the mobile operating system market share [40]. In this regard we can argue that the number of studies that use iOS is relatively low. What are the consequences of this in terms of understanding the performance of mobile Web apps on the iOS platform? It is possible that it prevents the optimization of mobile Web apps on the iOS platform since most optimization techniques are based on studies that did their experiments on an Android based device. It might be possible that these findings do not carry over, or not carry over in the same way to the iOS platform, thus resulting in the wrong aspects of Web apps being optimized. We are therefore in favor of doing experiments using iOS based devices with a focus on how the results differ from their Android counterparts. A recent paper investigating the trends and challenges of the mobile software engineering domain mirrors our findings regarding the high amount of scientific contributions featuring only the Android ecosystem [41]. As the mobile space is extremely dynamic they question the fate of the large body of Android-specific knowledge. Consequently they suggest the research community to focus more on the fundamental challenges of the mobile software engineering ecosystem so research results are more relevant and future-proof. As a first step they ask researchers to report on the generality of their conducted studies. We agree and support this advice.

Concerning the distribution of mobile browsers, given Safari's relative high percentage, 25.61% to be exact as of October 2021, of mobile browser market share worldwide the number of measurement-based studies is relative low [42]. However, we have already touched upon this issue above as its caused by the low number of studies using iOS based devices in our set of primary studies.

5.3. Insights and recommendations about the subjects (RQ3)

Regarding the Web apps used in the experiments, we see that there is a **strong dominance of the Alexa list** when it comes to a source to select sites from. The Alexa list provides the most popular websites worldwide and thus essentially shows the most visited sites for the average user and in that regard can be considered a solid choice as its representative of the browser behavior of must users. However, when selecting websites from the top of the list researchers should be aware that **typically the popular top 200 Web pages on Alexa tend to be *highly optimized***. The performance of these Web pages may not be typical and therefore not generalizable. Some papers acknowledge this possible bias and try to mitigate the problem. For example, S5 selects websites from various positions in the list, i.e., 30% from the pages from the bottom of Alea's 1 million websites. Apart from the possible lack of generalizability, there is research that shows that Alexa rankings can be manipulated and change significantly on a daily basis [14]. To combat these shortcomings, the Tranco list has been recently-introduced. The Tranco list is based on the combination of four existing lists (*i.e.,* Alexa, Umbrella, Quantcast, and Majestic). The Tranco list in allows researchers to filter out undesirable (*e.g.,* unavailable or malicious) domains, it is stable over time, and it has been designed for reducing the effort in replicating studies based upon it [14,27]. We suggest researchers to be aware of the aforementioned initiatives and act accordingly. Finally, **there were a number of primary studies where the Web apps were selected without any proper reasoning and a relative high number of papers that do not provide the actual URLs when using real websites**. This is not desirable as it negatively impacts the replicability of those studies.

5.4. Insights and recommendations about the platform (RQ4)

The distribution of the scope of the experiments show that most experiments focus on *page load only*, while in practice real users interact with mobile websites through gestures and other interactions. **Experiments based on page load only may not be representative of the actual QoE perceived by the users**, especially if we consider the high amount of JavaScript-based techniques used today for performance improvement, *e.g.,* lazy loading resources.

Finally, our extracted data reveals that a **relative large number of papers is not mentioning whether caching is enabled or disabled**

[11] https://ropensci.github.io/reproducibility-guide/sections/checklist

during the experiment. This is unfortunate as it makes it difficult to replicate the study as well as understanding how the results should be interpreted. This can also be related to the large number of papers not providing a replication package. There seems to be a general lack of transparency when describing the experimental environment.

5.5. Emerging challenges

By examining and analyzing the 'limitations', 'future work' and 'treats to validity' sections of our primary studies we found several overarching themes in terms of challenges encountered when performing measurement-based experiments on the mobile web. More specifically, three main challenges for researchers emerged:

- *low generalizability of obtained results* (17 occurrences). Often this is because of the limited number of hardware devices and websites used;
- *representativeness of the used metrics* (8 occurrences). In most cases this concerns the use of the page load time (PLT) metric. Overall authors find that better metrics are available, but these metrics can possibly influence the measurements or are significantly more difficult to measure;
- *measurement errors* (6 occurrences). This could be caused because of technical limitations of the measurement infrastructure; for example, separating the energy consumed by different processes is a well-known challenge for researchers [43].

The issue of low generalizability is something we have touched upon multiple times above. We suggest that future research should try to employ a more diverse selection of devices and websites to improve the generalizability of this field of research. Concerning the representativeness of used metrics, the rise of new widely-used tools like Google Lighthouse may improve this situation as we expect that browser vendors will be more and more inclined to cater to developers. Similarly, the issue of possible measurement errors may be mitigated by the development of new tools and models, as well as the usage of existing tools such as Android Runner, which supports researchers in following the empirical best practices by design [39].

6. Threats to validity

This section reports on the potential threats to validity of this study according to the Cook and Campbell categorization [8].

6.1. Internal validity

Internal threats to validity refer to the influence that extraneous variables may have on the design of the study [8]. This threat has been mitigated as much as possible by defining *a priori* and following a strict plan of the study, elaborated on in Section 2. This study plan has been iteratively defined by discussing it after each iteration among all the co-authors of this study. Moreover, we iteratively defined the classification framework by rigorously applying the keywording process. The synthesis of the collected data has been performed by applying basic well-known descriptive statistics. Also, during the horizontal analysis we made sanity checks of the extracted data by cross-analyzing parameters of the classification framework. The sanity checks have been done for each of the identified 15 potentially-relevant pairs of parameters of the classification framework and confirmed that *all* extracted data was consistent across the considered parameters. Below we report some representative examples of the performed sanity checks: there shall not be any study on energy consumption executed on an emulated device (since at the time of writing there is no accurate energy model for emulated devices), there shall not be any study targeting Safari on Android (since Safari is available only on iOS devices), all studies using the Monsoon power monitor collect at least one energy-related metric (otherwise it would not make any sense to use a power monitor in the experiment), etc. Finally, goal, research questions, and the search and selection phase of this study have been already reviewed by three independent researchers in the context of the initial version of this study [5], presented at the International Conference on Evaluation and Assessment in Software Engineering.

6.2. External validity

External threats to validity refer to the generalizability of the obtained results [8]. To mitigate this possible type of threats, we employed an academic search engine as well as used backward/forward snowballing. In addition, the inclusion and exclusion criteria were thoroughly discussed and defined collaboratively among all co-authors of this study; in this context the goal was to obtain a set of selection criteria which is minimal, but complete with respect to the goal of our study. Another potential threat may be the exclusion of papers not written in the English language. However, we deem this treat to be minimal as English is considered to be the main language of Science [44]. Finally, the focus on peer-reviewed papers only could be seen to be a risk but is actually intrinsic to our study design since we aim to focus exclusively on the state of the art presented in high-quality scientific studies. This potential bias did not impact our study significantly since considered papers have undergone a rigorous peer-review process, which is a well-established requirement for high quality publications.

6.3. Construct validity

Construct threats to validity refer to the selection of the primary studies with respect to the targeted goal and research questions [8]. A potential threat to the construct validity of this study might be due to the used search string; the used search string might not cover all or not return a representative set of studies on measurement-based experiments on the mobile Web. To mitigate the problem of primary studies not being able to properly answer the chosen research questions we performed an automatic search on all data sources indexed by Google Scholar. This accounts for potential biases due to publishers policies and business concerns. In addition, the search string was kept as general as possible to enable a high level of inclusiveness. Moreover, we further mitigated this potential threat by (i) piloting the search string against a set of known studies and (ii) by complementing it via the backward/-forward snowballing procedure [16]. Specifically, we identified the following studies for piloting our search string: [10,45,46], and [13]. Those papers have been selected as pilots based on our experience in the area (as a research group we are conducting measurement-based studies on the mobile Web since more than seven years). During the piloting of the search string, we (i) included three additional terms which allowed us to cover a higher number of potentially-relevant studies (i.e., "assessment", "analysis", "browser"), (ii) we removed keywords which were too generic for our search and were leading to too many false positives (i.e., "study", "investigation"), (iii) added the present participle form of the verbs related to the terms "assessment" ("assessing"), "analysis" ("analysing"), and "measurement" ("measuring"), and (iv) came to the conclusion that the snowballing procedure was necessary for mitigating the risk of having a high number of false negatives in our search due to the intrinsic rigidity of a keyword-based search. After collecting all relevant studies, we manually carried out the selection process using the chosen inclusion and exclusion criteria as discussed in Section II-B2.

6.4. Conclusion validity

Conclusion threats to validity refer to the relationship between the extracted data and the obtained results [8]. Potential biases during the data extraction process were mitigated by discussing the defined classification framework among all co-authors of this study. This way we could guarantee that the data extraction process was aligned with our targeted goal and research questions. Furthermore, we applied well-known best practices on the conduction of secondary studies during each phase of our study [8,17,47].

Finally, a complete replication package is publicly available, which allows third-party researchers to carry out independent replication and verification of our study. The replication package includes the raw data, scripts, and annotations produced during each phase of the study.

7. Conclusions

We conducted a systematic mapping study on measurement-based experiments on the mobile web. Starting from an initial selection of 786 potentially relevant papers, we followed a rigorous selection procedure to reduce them to a set of 33 primary studies. Each was analyzed by applying a predefined classification framework to answer our research questions.

The main findings of this study can be summarized as follows:

1. Performance and energy consumption are the most commonly considered aspects in measurement-based experiments on the mobile Web. Frequently used metrics to report these measurements are the Page Load Time, the Speed Index, and Time-to-Interactive for performance and Joules for energy consumption. To measure energy consumption primarily software-based tools are used.
2. All studies analyze their data using descriptive statistics; however, only a limited number of papers use hypothesis testing to statistically support their findings. In particular, studies that do not focus on performance or energy consumption rely exclusively on descriptive statistics.
3. A limited number of studies provide a replication package, although the ratio of studies that provide one has increased in recent years.
4. The most used device type on which to carry out experiments are smartphones running the Android operating system using the Google Chrome browser.
5. Most experiments use real websites that are selected through the Alexa list and are hosted on their original servers.
6. Most experiments focus on page load only, with caching disabled and using a non-simulated WiFi network.
7. Synthetic experimental subjects are more frequently used than real subjects in studies focusing on memory, energy, or networking.
8. Studies employing real devices tend to employ a more limited number of subjects, likely due to the difficulties related to scaling the execution of the experiments on real devices.
9. Studies employing synthetic subjects tend to employ a more limited number of subjects, likely due to the difficulty of crafting a large set of subjects automatically.

The obtained insights can help practitioners and researchers by providing an evidence-based overview of the state of the art on the common techniques, empirical practices, and tools used to perform measurement-based experiments targeting the mobile Web. In addition, we highlighted some challenges commonly experienced while performing measurement-based experiments on the mobile Web. Finding solutions for these challenges is an open research direction.

Appendix. Primary studies

See Table A.1.

Table A.1
Primary studies.

ID	Title	Authors	Year
S1	Mobile Web Browsing Under Memory Pressure [31]	Qazi, Ihsan Ayyub and Qazi, Zafar Ayyub and Benson, Theophilus A and Murtaza, Ghulam and Latif, Ehsan and Manan, Abdul and Tariq, Abrar	2020
S2	Web Browser Workload Characterization for Power Management on HMP Platforms [31]	Peters, Nadja and Park, Sangyoung and Chakraborty, Samarjit and Meurer, Benedikt and Payer, Hannes and Clifford, Daniel	2016
S3	Privacy as a proxy for Green Web browsing: Methodology and experimentation [48]	D'Ambrosio, Salvatore and De Pasquale, Salvatore and Iannone, Gerardo and Malandrino, Delfina and Negro, Alberto and Patimo, Giovanni and Scarano, Vittorio and Spinelli, Raffaele and Zaccagnino, Rocco	2017
S4 A	An empirical study of power consumption of Web-based communications in mobile phones [31]	Ayala, Inmaculada and Amor, Mercedes and Fuentes, Lidia and Munoz, Daniel	2017
S4B	An Energy Efficiency Study of Web-Based Communication in Android Phones [49]	Ayala, Inmaculada and Amor, Mercedes and Fuentes, Lidia	2019
S5	Investigating the Correlation between Performance Scores and Energy Consumption of Mobile Web Apps [50]	Chan-Jong-Chu, Kwame and Islam, Tanjina and Exposito, Miguel Morales and Sheombar, Sanjay and Valladares, Christian and Philippot, Olivier and Grua, Eoin Martino and Malavolta, Ivano	2020
S6	Why are Web Browsers Slow on Smartphones? [51]	Wang, Zhen and Lin, Felix Xiaozhu and Zhong, Lin and Chishtie, Mansoor	2015
S7	From 6.2 to 0.15 s - an Industrial Case Study on Mobile Web Performance [52]	van Riet, Jasper and Paganelli, Flavia and Malavolta, Ivano	2020
S8	The Web for Underpowered Mobile Devices: Lessons Learned from Google Glass [53]	Chauhan, Jagmohan and Kaafar, Mohamed Ali and Mahanti, Anirban	2015
S9 A	Demystifying the Imperfect Client-Side Cache Performance of Mobile Web Browsing [54]	Liu, Xuanzhe and Ma, Yun and Liu, Yunxin and Xie, Tao and Huang, Gang	2015
S9B	Characterizing Cache Usage for Mobile Web Applications [55]	Ma, Yun and Lu, Xuan and Zhang, Shuhui and Liu, Xuanzhe	2014
S9C	Measurement and Analysis of Mobile Web Cache Performance [56]	Ma, Yun and Liu, Xuanzhe and Zhang, Shuhui and Xiang, Ruirui and Liu, Yunxin and Xie, Tao	2015

(continued on next page)

Table A.1 (*continued*).

ID	Title	Authors	Year
S10	Performance Analysis of Web-browsing Speed in Smart Mobile Devices [57]	Kim, Yu-Doo and Moon, Il-Young	2013
S11	What is wrecking your dataplan? A measurement study of mobile web overhead [58]	Mendoza, Abner and Singh, Kapil and Gu, Guofei	2015
S12	Understanding Quality of Experiences on Different Mobile Browsers [59]	Tian, Deyu and Ma, Yun	2019
S13	Push or Request: An Investigation of HTTP/2 Server Push for Improving Mobile Performance [60]	Rosen, Sanae and Han, Bo and Hao, Shuai and Mao, Z Morley and Qian, Feng	2017
S14	mBenchLab: Measuring QoE of Web Applications using mobile devices [61]	Cecchet, Emmanuel and Sims, Robert and He, Xin and Shenoy, Prashant	2013
S15	Experimental Study of Energy and Bandwidth Costs of Web Advertisements on Smartphones [62]	Albasir, Abdurhman and Naik, Kshirasagar and Plourde, Bernard and Goel, Nishith	2014
S16	AMP up your Mobile Web Experience: Characterizing the Impact of Google's Accelerated Mobile Project [63]	Jun, Byungjin and Bustamante, Fabián E and Whang, Sung Yoon and Bischof, Zachary S	2019
S17	Analysis of the Energy Consumption of JavaScript Based Mobile Web Applications [64]	Miettinen, Antti P and Nurminen, Jukka K	2010
S18	Characterizing Resource Usage for Mobile Web Browsing [65]	Qian, Feng and Sen, Subhabrata and Spatscheck, Oliver	2014
S19	Refining Mobile Web Design for Reducing Energy Consumption of Mobile Terminals [66]	Ihara, Takuya and Doki, Suguru and Ogishi, Tomohiko and Tang, Suhua and Obana, Sadao	2015
S20	Need for Mobile Speed: A Historical Analysis of Mobile Web Performance [67]	Nejati, Javad and Luo, Meng and Nikiforakis, Nick and Balasubramanian, Aruna	2020
S21	Caching Does not Improve Mobile Web Performance (Much) [12]	Vesuna, Jamshed and Scott, Colin and Buettner, Michael and Piatek, Michael and Krishnamurthy, Arvind and Shenker, Scott	2016
S22	Web Experience in Mobile Networks: Lessons from Two Million Page Visits [68]	Rajiullah, Mohammad and Lutu, Andra and Khatouni, Ali Safari and Fida, Mah-Rukh and Mellia, Marco and Brunstrom, Anna and Alay, Ozgu and Alfredsson, Stefan and Mancuso, Vincenzo	2019
S23	Assessing the Impact of Service Workers on the Energy Efficiency of Progressive Web Apps [13]	Malavolta, Ivano and Procaccianti, Giuseppe and Noorland, Paul and Vukmirovic, Petar	2017
S24	Evaluating the Impact of Caching on the Energy Consumption and Performance of Progressive Web Apps [69]	Malavolta, Ivano and Chinnappan, Katerina and Jasmontas, Lukas and Gupta, Sarthak and Soltany, Kaveh Ali Karam	2020
S25	Who killed my battery: Analyzing mobile browser energy consumption [46]	Thiagarajan, Narendran and Aggarwal, Gaurav and Nicoara, Angela and Boneh, Dan and Singh, Jatinder Pal	2012
S26	An In-depth study of Mobile Browser Performance [45]	Nejati, Javad and Balasubramanian, Aruna	2016
S27	Measuring AJAX Performance on a GPRS Mobile Platform [70]	Xie, Feng and Parsons, David	2008
S28	Evaluation of Techniques for web 3D Graphics Animation on Portable Devices [71]	Kapetanakis, Kostas and Panagiotakis, Spyros	2012
S29	WebMythBusters: An In-depth Study of Mobile Web Experience [72]	Park, Seonghoon and Choi, Yonghun and Cha, Hojung	2021
S30	Characterizing Embedded Web Browsing in Mobile Apps [73]	Tian, Deyu and Ma, Yun and Balasubramanian, Aruna and Liu, Yunxin and Huang, Gang and Liu, Xuanzhe	2021
S31 A	A Tale of Two Fashions: An Empirical Study on the Performance of Native Apps and Web Apps on Android [74]	Ma, Yun and Liu, Xuanzhe and Liu, Yi and Liu, Yunxin and Huang, Gang	2018
S31B	Characterizing restful web services usage on smartphones: A tale of native apps and web apps [75]	Liu, Yi and Liu, Xuanzhe and Ma, Yun and Liu, Yunxin and Zheng, Zibin and Huang, Gang and Blake, M Brian	2015
S32	WebMedic: Disentangling the Memory-Functionality Tension for the Next Billion Mobile Web Users [76]	Naseer, Usama and Benson, Theophilus A and Netravali, Ravi	2021
S33	PWA vs the Others: A Comparative Study on the UI Energy-Efficiency of Progressive Web Apps [77]	Huber, Stefan and Demetz, Lukas and Felderer, Michael	2021

References

[1] StatCounter, Desktop vs Mobile vs Tablet Market Share Worldwide Nov 2015 - Nov 2020, URL https://gs.statcounter.com/platform-market-share/desktop-mobile-tablet/worldwide/#monthly-201511-202011, Accessed: 2020-12-15.

[2] Philippe De Ryck, Lieven Desmet, Frank Piessens, Martin Johns, The browser as a platform, in: Primer on Client-Side Web Security, Springer, 2014, pp. 25–32.

[3] Kit Eaton, How one second could cost amazon $1.6 billion in sales, Fast Company 14 (2012).

[4] Ivano Malavolta, Eoin Martino Grua, Cheng-Yu Lam, Randy de Vries, Franky Tan, Eric Zielinski, Michael Peters, Luuk Kaandorp, A framework for the automatic execution of measurement-based experiments on android devices, in: 35th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW '20), ACM, 2020, pp. 61–66, URL http://www.ivanomalavolta.com/files/papers/A_Mobile_2020.pdf.

[5] Omar de Munk, Ivano Malavolta, Measurement-based experiments on the mobile web: A systematic mapping study, in: Proceedings of the International Conference on Evaluation and Assessment on Software Engineering (EASE), ACM, 2021, pp. 191–200, URL http://www.ivanomalavolta.com/files/papers/EASE_2021.pdf.

[6] K. Peterson, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, 2015.

[7] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, 2013.

[8] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, Anders Wesslén, Experimentation in Software Engineering, Springer Science & Business Media, 2012.

[9] Victor R. Basili, Gianluigi Caldiera, H. Dieter Rombach, The goal question metric approach, in: Encyclopedia of Software Engineering, Vol. 2, Wiley, 1994, pp. 528–532.

[10] Ivano Malavolta, Katerina Chinnappan, Lukas Jasmontas, Sarthak Gupta, Kaveh Ali Karam Soltany, Evaluating the impact of caching on the energy consumption and performance of progressive web apps, in: Proceedings of the IEEE/ACM 7th International Conference on Mobile Software Engineering and Systems, 2020, pp. 109–119, http://dx.doi.org/10.1145/3387905.3388593, URL http://www.ivanomalavolta.com/files/papers/MOBILESoft_Caching_PWA_2020.pdf.

[11] Jasper van Riet, Flavia Paganelli, Ivano Malavolta, From 6.2 to 0.15 seconds - an industrial case study on mobile web performance, in: 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME), 2020, pp. 746–755, http://dx.doi.org/10.1109/ICSME46990.2020.00084.

[12] Jamshed Vesuna, Colin Scott, Michael Buettner, Michael Piatek, Arvind Krishnamurthy, Scott Shenker, Caching doesn't improve mobile web performance (much), in: 2016 USENIX Annual Technical Conference (USENIX/ATC 16), 2016, pp. 159–165.

[13] Ivano Malavolta, Giuseppe Procaccianti, Paul Noorland, Petar Vukmirovic, Assessing the impact of service workers on the energy efficiency of progressive web apps, in: Proceedings of the International Conference on Mobile Software Engineering and Systems, MOBILESoft '17, Buenos Aires, Argentina, May, 2017, 2017, pp. 35–45, http://dx.doi.org/10.1109/MOBILESoft.2017.7, URL http://www.ivanomalavolta.com/files/papers/Mobilesoft_2017.pdf.

[14] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, Wouter Joosen, Tranco: A research-oriented top sites ranking hardened against manipulation, 2018, arXiv preprint arXiv:1806.01156.

[15] Michael Gusenbauer, Google scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases, Scientometrics 118 (1) (2019) 177–214.

[16] Claes Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–10.

[17] Kai Petersen, Sairam Vakkalanka, Ludwik Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Inf. Softw. Technol. 64 (2015) 1–18.

[18] Kai Petersen, Robert Feldt, Shahid Mujtaba, Michael Mattsson, Systematic mapping studies in software engineering, in: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, in: EASE'08, British Computer Society, Swinton, UK, UK, 2008, pp. 68–77.

[19] Paolo Di Francesco, Patricia Lago, Ivano Malavolta, Architecting with microservices: A systematic mapping study, J. Syst. Softw. 150 (2019) 77–97, http://dx.doi.org/10.1016/j.jss.2019.01.001, URL http://www.ivanomalavolta.com/files/papers/JSS_MSA_2019.pdf.

[20] Mirco Franzago, Davide Di Ruscio, Ivano Malavolta, Henry Muccini, Collaborative model-driven software engineering: a classification framework and a research map, IEEE Trans. Softw. Eng. 14 (12) (2018) 1146–1175, http://dx.doi.org/10.1109/TSE.2017.2755039, URL http://www.ivanomalavolta.com/files/papers/TSE_2018.pdf.

[21] Barbara A. Kitchenham, Stuart Charters, Guidelines for performing systematic literature reviews in software engineering, Technical Report EBSE-2007-01, Keele University and University of Durham, 2007.

[22] Giuseppina Casalaro, Giulio Cattivera, Federico Ciccozzi, Ivano Malavolta, Andreas Wortmann, Patrizio Pelliccione, Model-driven engineering for mobile robotic systems: a systematic mapping study, Software and Systems Modeling (2021) 1–31.

[23] Daniela S. Cruzes, Tore Dybå, Research synthesis in software engineering: A tertiary study, Inf. Softw. Technol. 53 (5) (2011) 440–455.

[24] Patrick E. McKnight, Julius Najab, Mann-whitney u test, Corsini Encycl. Psychol. (2010) 1.

[25] Carlo Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, in: Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commericiali Di Firenze, 8, 1936, pp. 3–62.

[26] Norman Cliff, Dominance statistics: Ordinal analyses to answer ordinal questions, Psychol. Bull. 114 (3) (1993) 494.

[27] Victor Le Pochat, Tom van Goethem, Wouter Joosen, Rigging research results by manipulating top websites rankings, 2018, CoRR arXiv:1806.01156.

[28] Luca Ardito, Giuseppe Procaccianti, Marco Torchiano, Giuseppe Migliore, Profiling power consumption on mobile devices.

[29] Gustavo Pinto, Fernando Castor, Energy efficiency: a new concern for application software developers, Commun. ACM 60 (12) (2017) 68–75.

[30] Jiri Hosek, Michal Ries, Pavel Vajsar, Lubos Nagy, Zdenek Sulc, Petr Hais, Radek Penizek, Mobile web QoE study for smartphones, in: 2013 IEEE Globecom Workshops (GC Wkshps), IEEE, 2013, pp. 1157–1161.

[31] Ihsan Ayyub Qazi, Zafar Ayyub Qazi, Theophilus A. Benson, Ghulam Murtaza, Ehsan Latif, Abdul Manan, Abrar Tariq, Mobile web browsing under memory pressure, ACM SIGCOMM Comput. Commun. Rev. 50 (4) (2020) 35–48.

[32] Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, Jennifer L. Tackett, Abandon statistical significance, Amer. Statist. 73 (sup1) (2019) 235–245.

[33] David Colquhoun, An investigation of the false discovery rate and the misinterpretation of p-values, Royal Soc. Open Sci. 1 (3) (2014) 140216.

[34] Gail M. Sullivan, Richard Feinn, Using effect size or why the P value is not enough, J. Graduate Med. Educ. 4 (3) (2012) 279.

[35] Victoria Stodden, The scientific method in practice: Reproducibility in the computational sciences, 2010.

[36] John E. Boylan, Reproducibility, IMA J. Manag. Math. 27 (2) (2016) 107–108.

[37] David B. Resnik, Adil E. Shamoo, Reproducibility and research integrity, Account. Res. 24 (2) (2017) 116–123.

[38] Erin D. Foster, Ariel Deardorff, Open science framework (OSF), J. Med. Lib. Assoc.: JMLA 105 (2) (2017) 203.

[39] Ivano Malavolta, Eoin Martino Grua, Cheng-Yu Lam, Randy de Vries, Franky Tan, Eric Zielinski, Michael Peters, Luuk Kaandorp, A framework for the automatic execution of measurement-based experiments on android devices, in: 35th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW-20), pp. 61–66.

[40] StatCounter, Mobile operating system market share worldwide nov 2020, 2020, Accessed: 2020-12-18, https://gs.statcounter.com/os-market-share/mobile/worldwide.

[41] Luciano Baresi, William G. Griswold, Grace A. Lewis, Marco Autili, Ivano Malavolta, Christine Julien, Trends and challenges for software engineering in the mobile domain, IEEE Softw. 38 (1) (2020) 88–96.

[42] StatCounter, Mobile browser market share worldwide nov 2020, 2020, Accessed: 2020-12-18, https://gs.statcounter.com/browser-market-share/mobile/worldwide.

[43] Rubén Saborido, Venera Venera Arnaoudova, Giovanni Beltrame, Foutse Khomh, Giuliano Antoniol, On the impact of sampling frequency on software energy measurements, Technical Report, PeerJ PrePrints, 2015.

[44] David G. Drubin, Douglas R. Kellogg, English as the universal language of science: opportunities and challenges, Mol. Biol. Cell 23 (8) (2012) 1399.

[45] Javad Nejati, Aruna Balasubramanian, An in-depth study of mobile browser performance, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 1305–1315.

[46] Narendran Thiagarajan, Gaurav Aggarwal, Angela Nicoara, Dan Boneh, Jatinder Pal Singh, Who killed my battery? Analyzing mobile browser energy consumption, in: Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 41–50.

[47] Barbara Kitchenham, Pearl Brereton, A systematic review of systematic review process research in software engineering, Inf. Softw. Technol. 55 (12) (2013) 2049–2075.

[48] Salvatore D'Ambrosio, Salvatore De Pasquale, Gerardo Iannone, Delfina Malandrino, Alberto Negro, Giovanni Patimo, Vittorio Scarano, Raffaele Spinelli, Rocco Zaccagnino, Privacy as a proxy for green web browsing: Methodology and experimentation, Comput. Netw. 126 (2017) 81–99.

[49] Inmaculada Ayala, Mercedes Amor, Lidia Fuentes, An energy efficiency study of web-based communication in android phones, Sci. Program. 2019 (2019).

[50] Kwame Chan-Jong-Chu, Tanjina Islam, Miguel Morales Exposito, Sanjay Sheombar, Christian Valladares, Olivier Philippot, Eoin Martino Grua, Ivano Malavolta, Investigating the correlation between performance scores and energy consumption of mobile web apps, in: Proceedings of the Evaluation and Assessment in Software Engineering, 2020, pp. 190–199.

[51] Zhen Wang, Felix Xiaozhu Lin, Lin Zhong, Mansoor Chishtie, Why are web browsers slow on smartphones?, in: Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, 2011, pp. 91–96.

[52] Jasper van Riet, Flavia Paganelli, Ivano Malavolta, From 6.2 to 0.15 seconds–an industrial case study on mobile web performance, in: 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, 2020, pp. 746–755.

[53] Jagmohan Chauhan, Mohamed Ali Kaafar, Anirban Mahanti, The web for underpowered mobile devices: Lessons learned from google glass, IEEE Internet Comput. 22 (3) (2018) 38–47.

[54] Xuanzhe Liu, Yun Ma, Yunxin Liu, Tao Xie, Gang Huang, Demystifying the imperfect client-side cache performance of mobile web browsing, IEEE Trans. Mob. Comput. 15 (9) (2015) 2206–2220.

[55] Yun Ma, Xuan Lu, Shuhui Zhang, Xuanzhe Liu, Characterizing cache usage for mobile web applications, in: Proceedings of the 6th Asia-Pacific Symposium on Internetware on Internetware, 2014, pp. 68–71.

[56] Yun Ma, Xuanzhe Liu, Shuhui Zhang, Ruirui Xiang, Yunxin Liu, Tao Xie, Measurement and analysis of mobile web cache performance, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 691–701.

[57] Yu-Doo Kim, Il-Young Moon, Performance analysis of web-browsing speed in smart mobile devices, Int. J. Smart Home 7 (2) (2013) 39–48.

[58] Abner Mendoza, Kapil Singh, Guofei Gu, What is wrecking your data plan? a measurement study of mobile web overhead, in: 2015 IEEE Conference on Computer Communications (INFOCOM), IEEE, 2015, pp. 2740–2748.

[59] Deyu Tian, Yun Ma, Understanding quality of experiences on different mobile browsers, in: Proceedings of the 11th Asia-Pacific Symposium on Internetware, 2019, pp. 1–10.

[60] Sanae Rosen, Bo Han, Shuai Hao, Z. Morley Mao, Feng Qian, Push or request: An investigation of HTTP/2 server push for improving mobile performance, in: Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 459–468.

[61] Emmanuel Cecchet, Robert Sims, Xin He, Prashant Shenoy, MBenchLab: Measuring QoE of web applications using mobile devices, in: 2013 IEEE/ACM 21st International Symposium on Quality of Service (IWQoS), IEEE, 2013, pp. 1–10.

[62] Abdurhman Albasir, Kshirasagar Naik, Bernard Plourde, Nishith Goel, Experimental study of energy and bandwidth costs of web advertisements on smartphones, in: 6th International Conference on Mobile Computing, Applications and Services, IEEE, 2014, pp. 90–97.

[63] Byungjin Jun, Fabián E. Bustamante, Sung Yoon Whang, Zachary S. Bischof, AMP up your mobile web experience: Characterizing the impact of google's accelerated mobile project, in: The 25th Annual International Conference on Mobile Computing and Networking, 2019, pp. 1–14.

[64] Antti P. Miettinen, Jukka K. Nurminen, Analysis of the energy consumption of javascript based mobile web applications, in: International Conference on Mobile Lightweight Wireless Systems, Springer, 2010, pp. 124–135.

[65] Feng Qian, Subhabrata Sen, Oliver Spatscheck, Characterizing resource usage for mobile web browsing, in: Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, 2014, pp. 218–231.

[66] Takuya Ihara, Suguru Doki, Tomohiko Ogishi, Suhua Tang, Sadao Obana, Refining mobile web design for reducing energy consumption of mobile terminals, in: 2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies, IEEE, 2015, pp. 13–18.

[67] Javad Nejati, Meng Luo, Nick Nikiforakis, Aruna Balasubramanian, Need for Mobile Speed: A Historical Analysis of Mobile Web Performance.

[68] Mohammad Rajiullah, Andra Lutu, Ali Safari Khatouni, Mah-Rukh Fida, Marco Mellia, Anna Brunstrom, Ozgu Alay, Stefan Alfredsson, Vincenzo Mancuso, Web experience in mobile networks: Lessons from two million page visits, in: The World Wide Web Conference, 2019, pp. 1532–1543.

[69] Ivano Malavolta, Katerina Chinnappan, Lukas Jasmontas, Sarthak Gupta, Kaveh Ali Karam Soltany, Evaluating the impact of caching on the energy consumption and performance of progressive web apps, in: 7th IEEE/ACM International Conference on Mobile Software Engineering and Systems 2020, 2020.

[70] Feng Xie, David Parsons, Measuring ajax performance on a GPRS mobile platform, in: 7th International Conference on Applications and Principles of Information Science (APIS2008), Citeseer, 2008, pp. 28–29.

[71] Kostas Kapetanakis, Spyros Panagiotakis, Evaluation of techniques for web 3d graphics animation on portable devices, in: 2012 International Conference on Telecommunications and Multimedia (TEMU), IEEE, 2012, pp. 152–157.

[72] Seonghoon Park, Yonghun Choi, Hojung Cha, Webmythbusters: An in-depth study of mobile web experience, in: IEEE INFOCOM 2021-IEEE Conference on Computer Communications, IEEE, 2021, pp. 1–10.

[73] Deyu Tian, Yun Ma, Aruna Balasubramanian, Yunxin Liu, Gang Huang, Xuanzhe Liu, Characterizing embedded web browsing in mobile apps, IEEE Trans. Mob. Comput. (2021).

[74] Yun Ma, Xuanzhe Liu, Yi Liu, Yunxin Liu, Gang Huang, A tale of two fashions: An empirical study on the performance of native apps and web apps on android, IEEE Trans. Mob. Comput. 17 (5) (2017) 990–1003.

[75] Yi Liu, Xuanzhe Liu, Yun Ma, Yunxin Liu, Zibin Zheng, Gang Huang, M. Brian Blake, Characterizing restful web services usage on smartphones: A tale of native apps and web apps, in: 2015 IEEE International Conference on Web Services, IEEE, 2015, pp. 337–344.

[76] Usama Naseer, Theophilus A. Benson, Ravi Netravali, WebMedic: Disentangling the memory-functionality tension for the next billion mobile web users, in: Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications, 2021, pp. 71–77.

[77] Stefan Huber, Lukas Demetz, Michael Felderer, PWA vs the others: A comparative study on the UI energy-efficiency of progressive web apps, in: International Conference on Web Engineering, Springer, 2021, pp. 464–479.